GEOProcessing 2014

The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services

ISBN: 978-1-61208-326-1

March 23 - 27, 2014

Barcelona, Spain

GEOProcessing 2014 Editors

Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany
Yerach Doytsher, Technion - Israel Institute of Technology, Haifa, Israel

GEOProcessing 2014

Foreword

The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2014), held between March 23-27, 2014 in Barcelona, Spain, brought together researchers from the academia and practitioners from the industry in order to address fundamentals of advances in geographic information systems and the new applications related to them using the Web Services. Such systems can be used for assessment, modeling and prognosis of emergencies. As an example, these systems can be used for assessment of accidents from chemical pollution by considering hazardous chemical zones dimensions represented on a computer map of the region's territory.

Geographical sensors and satellites provide a huge volume of spatial data which is available on the Web. Making use of Web Services, the users are able to provision and use these services instead of only performing document searching. These services are published in a directory and may be automatically discovered in a given context by software agents. Accessing large digital geographical libraries with geo-spatial information raises some challenges with respect to data semantics, interfaces, data accuracy and updates, distributed processing, as well as with discovery, indexing and integration of geographical information systems; this raise the issue of distributed catalogs forming a federation of spatial databases. Some spatial data infrastructures use service-oriented architecture for accessing these large databases via Web Services.

We take here the opportunity to warmly thank all the members of the GEOProcessing 2014 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to GEOProcessing 2014. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the GEOProcessing 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that GEOProcessing 2014 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the areas of geographic information systems, applications and services.

We are convinced that the participants found the event useful and communications very open. We hope that Barcelona, Spain, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

GEOProcessing 2014 Chairs:

Monica De Martino, Consiglio Nazionale delle Ricerche - Genova, Italy
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany
Bernd Resch, Massachusetts Institute of Technology - Cambridge, USA
Yerach Doytsher, Technion - Israel Institute of Technology, Haifa, Israel

GEOProcessing 2014

COMMITTEE

GEOProcessing Advisory Chairs

Monica De Martino, Consiglio Nazionale delle Ricerche - Genova, Italy
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany
Bernd Resch, Massachusetts Institute of Technology - Cambridge, USA
Yerach Doytsher, Technion - Israel Institute of Technology, Haifa, Israel

GEOProcessing 2014 Technical Program Committee

Diana F. Adamatti, Universidade Federal do Rio Grande, Brazil
Ayman Al-Serafi, Teradata Corporation, USA
Mirko Albani, European Space Agency, Italy
Riccardo Albertoni, IMATI-CNR, Italy
Francesc Antón Castro, Denmark´s National Space Institute, Denmark
Thierry Badard, Université Laval - Québec, Canada
Petko Bakalov, Environmental Systems Research Institute, USA
Fabiano Baldo, Santa Catarina State University, Brazil
Fabian D. Barbato, ORT University - Montevideo, Uruguay
Thomas Barkowsky, University of Bremen, Germany
Reinaldo Bezerra Braga, Federal University of Ceará, Brazil
Budhendra L. Bhaduri, Oak Ridge National Laboratory, USA
Ling Bian, University at Buffalo, USA
Sandro Bimonte, Irstea | TSCF - Clermont Ferrand, France
Stefano Borgo, Institute of Cognitive Sciences and Technologies - CNR, Italy
Giuseppe Borruso, University of Trieste, Italy
Boyan Brodaric, Geological Survey of Canada, Canada
Jean Brodeur, Natural Resources Canada / Government of Canada, Canada
David Brosset, Naval Academy Research Institute, France
Michael Cathcart, Electro-Optical Systems Laboratory / GTRI Georgia Institute of Technology, USA
Mete Celik, Erciyes University, Turkey
Xin Chen, NAVTEQ Corporation - Chicago, USA
Chi-Yin Chow, City University of Hong Kong, Hong Kong
Christophe Claramunt Naval Academy Research Institute, France
Konstantin Clemens, TU-Berlin, Germany
Eliseo Clementini, University of L’Aquila, Italy
Ana Cristina Costa, Universidade Nova de Lisboa, Portugal
Chenyun Dai, Purdue University, USA
Monica De Martino, Consiglio Nazionale delle Ricerche (CNR) - Genova, Italy
Anselmo C. de Paiva, Universidade Federal do Maranhão, Brazil
Cláudio de Souza Baptista, University of Campina Grande, Brazil
Yerach Doytsher, Technion - Israel Institute of Technology, Haifa, Israel
Suzana Dragicevic, Simon Fraser University- Burnaby, Canada

Alenka Poplin, HafenCity University Hamburg, Germany
David Prosperi, Florida Atlantic University, USA
Sigrid Reiter, University of Liège, Belgium
Matthias Renz, Ludwig-Maximilians Universität München, Germany
Bernd Resch, Massachusetts Institute of Technology/Senseable City Lab - Cambridge, USA
Kai-Florian Richter, Department of Geography - University of Zurich, Switzerland
Henry Roig Llacer, Institute of Geosciences - University of Brasilia, Brazil
Sergio Rosim, National Institute for Space Research, Brazil
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany
Markus Schneider, University of Florida, USA
Shashi Shekhar, University of Minnesota, USA
Spiros Skiadopoulos, University of Peloponnese - Tripoli, Hellas
Frank Steinicke, Institut für Mensch-Computer-Medien & Institut für Informatik - Würzburg, Germany
Lena Strömbäck, SMHI, Sweden
Kazutoshi Sumiya, University of Hyogo, Japan
Juergen Symanzik, Utah State University - Logan, USA
Ali Tahir, University College Dublin, Ireland
Naohisa Takahashi, Nagoya Institute of Technology, Japan
Ergin Tari, Istanbul Technical University, Turkey
Maristela Terto de Holanda, University of Brasilia, Brazil
Jean-Claude Thill, University of North Carolina at Charlotte, USA
Luigi Troiano, University of Sannio, Italy
Theodore Tsiligiridis, Agricultural University of Athens, Greece
E. Lynn Usery, U.S. Geological Survey - Rolla, USA
Iván Esteban Villalón Turrubiates, Universidad Jesuita de Guadalajara, México
Jue Wang, Washington University in St. Louis, USA
Iris Weber, Institut für Planetologie, Westfälische Wilhelms-Universität Münster, Germany
Nancy Wiegand, University of Wisconsin-Madison, USA
Eric B. Wolf, US Geological Survey - Boulder, USA
Ouri Wolfson, University of Illinois - Chicago, USA
Mike Worboys, University of Maine - Orono, USA
Ningchuan Xiao, The Ohio State University - Columbus, USA
Zhangcai Yin, Wuhan University of Technology, China
May Yuan, Center for Spatial Analysis and Geoinformatics Program, College of Atmospheric and Geographic Sciences, University of Oklahoma, USA
Karine Zeitouni, University of Versailles Saint-Quentin, France
Chuanrong Zhang, University of Connecticut - Storrs, USA
Wenbing Zhao, Cleveland State University, USA

Copyright Information

Table of Contents

Data Collection Architecture for Field Research in Heterogeneous Computational Environments

Henrique P. de F. Filho, Maristela Holanda, Bernardo Macêdo, Renata Nunes, Paulo Brener
Department of Computer Science
University of Brasilia
Brasilia, Brazil
henripff@gmail.com, mholanda@cic.unb.br, bernardo-macedo@hotmail.com, cmn.renata@gmail.com, paulo_brener_12@hotmail.com

Henrique Llacer Roig
Institute of Geoscience
University of Brasilia
Brasilia, Brazil
roig@unb.br

Abstract—**The computational environment based on wireless communication made it possible to access information anywhere and at anytime, which is favorable for data collection in field research. Currently, most architecture for data collection are designed for a specific purpose, using specific technologies that are limited regarding data, network, synchronization and device type. This paper presents an architecture for field research data collection that works in heterogeneous computational environments and supports vector geographic data. This architecture was validated with a case study conducted at the Institute of Geosciences (IG) of the University of Brasilia (UNB).**

Keywords-Data collection; field research; architecture; heterogeneous computational environments; vector geographic data

I. INTRODUCTION

The technological advances of the last decades associated with wireless technology communication are intended to allow access to information anywhere and at anytime, making a favorable environment for the development of computational systems, which are aiming at collecting data in field research [16].

One of the challenges in this environment is to assure the consistent state of the database involved in the system. There are databases in mobile devices and the central database [21].

Nowadays, most of the data collecting architectures work in homogeneous computational environments [16][18][21]. In other words, they use specific technologies concerning, for instance, data, network, synchronization, and device types, and they fulfill a specific data gathering need.

In this context, this paper proposes an architecture for collecting data in a heterogeneous computational environment where information stored on each mobile device is replicated and correctly integrated into a central database. The proposed architecture not only collects conventional data, but also vector geographic data, because the demand for this type of data is high in geology and geoscience research, and spatial remote sensing.

This paper is organized as follows: Section II presents other work related to this research study; Section III deals with the main aspects of a system for mobile data collection; Section IV presents the proposed data collection architecture; Section V presents the case study; Section VI presents the conclusions.

II. RELATED WORK

A software suite that can be used for data collection in field research inserted in a mobile computational environment is available on the market. These software have architecture for data collection that caters to diverse contexts of field research, however, they are tied to specific technologies, which is a limitation of their use. Nokia Data Gathering solution and AuditMagic are examples of these software applications [3][16][18].

Nokia Data Gathering is a system that allows the collection of data on mobile devices and the transmission of results to real-time analysis, in accordance with the access to a network data communication. The system allows the creation and delivery of questionnaires for mobile phones and integration into a database using a pre-existing cellular network common [16][18]. With regards to this solution, a problem was identified: it is a closed architecture tool that only works for Nokia cell phone devices.

AuditMatic has a set of solutions for developing instruments to collect field data associated analysis tools. However, this tool is also a package deal where you cannot make changes according to the needs of each research [3][16].

Most architectures for data collection in field research proposed by researchers meet only a single context of data collection, namely, the context in which the research is embedded. The goal of this paper is to propose an architecture that meets the needs of data collection of this research. An example is the architecture for collection and dissemination of weather data in the state of Piauí proposed by researchers from EMBRAPA (Brazilian Agricultural Research Corporation) and the Faculty of Technological Education of Teresina [21].

Another example is the system architecture for data collection in a mobile computing environment where Internet access is intermittent, which was validated in the research of Rural School Transportation, developed by the Centro

Interdisciplinar de Estudos em Transportes (CEFTRU/UNB) in partnership with the Fundo Nacional de Desenvolvimento da Educação (FNDE), whose objective is to evaluate the reality of school transport in rural Brazil. [16] This architecture has mechanisms for fault tolerance; however, it does not support geographic data and uses a specific technology that meets only the context of the collection in question.

As previously stated, these architectures serve only these context specific collections using specifics technologies, specifics synchronization protocols that best fit the types of data to be synchronized, whether architectures can be used in other context collections or not.

The architectures proposed by researchers, which have the same objective as the architecture proposed in this paper, do not support the collection of geographic data.

Ji [23] presents an architecture that aims to collect georeferenced data. The communication between client and server is done via web Services using the technology of wireless networks.

III. MOBILE DATA COLLECTING SYSTEM

The Mobile Data Collection System (MDCS) allows the collection of remote geographic locations and the transmission of data to central locations - data storing repositories through a wireless network. It is a combination of a client application running on mobile devices, wireless network infrastructure and remotely accessible database servers [11].

Most of these systems share the same principles and guidelines for remote data collection. As shown in Figure 1, the process begins by developing the application with a form, which contains a set of questions to collect relevant data. Data collectors use these forms on the mobile device to collect real data in the field. It is possible to upload the data to a database in the central server [11].

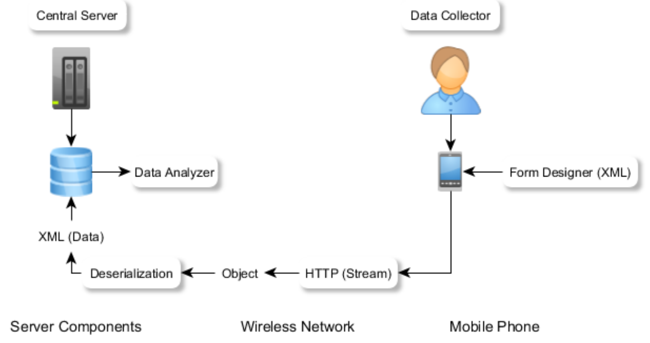


Figure 1. Flow diagram of a MDCS. Adapted from [11].

One challenge in MDCS is the data synchronization. In mobile environments, synchronization may be defined as the act of establishing equivalence between collections of clients and server databases [4][6]. For this synchronization to occur data replication is necessary.

Replication is the process in which transactions executed in a database are propagated synchronously or asynchronously to one or more databases in a serial manner, i.e., this means that all transactions are replicated in the same order in which they were requested [20]. If this does not happen, there may be inconsistencies between the mobile unit and the server in a mobile environment. Spreading asynchronously refers to a way of storing and sending replication, i.e., the operations to be replicated from mobile units can be stored in a local database until they are propagated to the server at the time of synchronization, since sometimes the connection is impossible [20].

The replication may be partial or total. In partial replication only part of the data from a database is replicated to other databases, while in full replication all data is replicated from one database to another [8]. With full replication, at least one copy is available, and the reliability is increased, since the user does not depend solely on the data available at a single location. In cases where a fault occurs in the system, data replication is requested by the user [14].

A synchronization protocol defines the workflow for communication over a section of data synchronization when the mobile device is connected to the fixed network. It should support the identification of records, protocol commands common to the local database and network synchronization, and be able to support the identification and resolution of synchronization conflicts [17].

IV. THE PROPOSED ARCHITECTURE

The goal of an architecture for data collection in field research that works on heterogeneous computational environments is to disassociate the specificities of each collection with the used technologies, providing an architecture that can be used in various contexts for data collection. Subsequently, in this study, for a data collection architecture to work in several different computational environments this architecture must take two aspects into account: interoperability and flexibility.

Interoperability is required so that architecture is not dependent on specific technology, but covering the possibility of using several technologies. Flexibility is required in two directions: the first to meet the diverse dimensions that data collection in the field can have, i.e., fulfill the data collection involving one or more institutions. A specific data collection can store the collected data in a single database located in a specific place or multiple databases located in several different places. The second is to make the necessary changes needed to pass the architecture to suit the contexts of data collection different, i.e., that the architecture is susceptible to changes - it can be configurable and reconfigurable.

An architecture for data collection in field research that supports vector geographic data must use a *Data Base Management System* (DBMS) with support for spatial data in both collection devices as the servers involved in collecting.

Abstractly, the proposed architecture consists of three stages: the collection field, synchronization and storage, as shown in Figure 2. Each of these stages is described as follows.

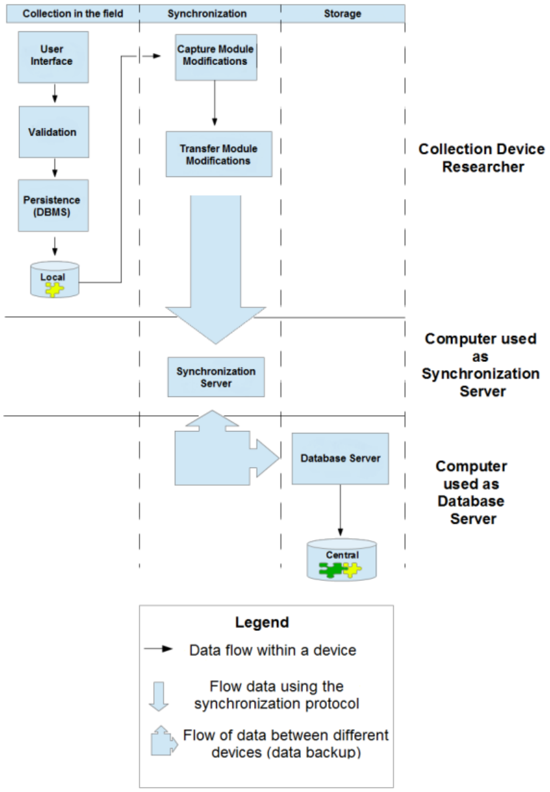

Collection in the field
Synchronization
Storage
User Interface
Validation
Persistence (DBMS)
Local
Capture Module Modifications
Transfer Module Modifications
Collection Device Researcher
Computer used as Synchronization Server
Synchronization Server
Database Server
Computer used as Database Server
Central
Legend
Data flow within a device
Flow data using the synchronization protocol
Flow of data between different devices (data backup)

Figure 2. Abstract architecture of the heterogeneous data collection system.

Data collection in the field occurs through a computer system in the researchers' mobile devices. This computer system must have a user interface, a method of data validation and a persistence layer. The user interface is responsible for entering data into the system, the validation method is responsible for validating the data entered so that relevant data are not forgotten or recorded incorrectly and the persistence layer is responsible for storing the data in the database mobile devices. The persistence layer must necessarily store the data in a DBMS that supports spatial data.

After collecting the data, there is a synchronization step, where the collected data is synchronized with the synchronization server. The objective of synchronization is to generate replicas, through a full replication of the collected information, to the server that synchronizes data with the database server involved in the core collection. This synchronization is specified in the next section. Finally, data servers are stored in the core data set involved in the collection via data backup.

The possibility of having more than one server central database is taken into consideration in this architecture, given the possible dimensions of data collection, contributing to the architecture's characteristic flexibility. When we have more than one server database replication, server data synchronization occurs for these servers, creating data redundancy, which prevents the architecture from having a single point of failure.

The synchronization protocol is used in architecture that will lay down the interoperability and flexibility of the architecture (flexibility towards the possibility of change, to adapt to contexts of different data collection), so the protocol for the architecture should be characterized as interoperability, and should be a free protocol, i.e., not proprietary protocol.

A. *Synchronization Stage*

The proposed synchronization is divided into two stages: capturing *Structured Query Language* (SQL) [8] statements and the transfer of these instructions to the synchronization server. The capture module modifications are responsible for capturing the SQL database from the mobile device in the event of any change in the database. The transfer module modification plays a client role in a client-server architecture, synchronizing the changes to the SQL statement capture module with the synchronization server.

1) *Capture Module Modifications*

The capture of these new SQL statements is by searching for data that have a flag with value 1. This flag determines if the data has been synchronized or not. If it is set as 1, then the data has not been synchronized yet; if it is set as 0, then the data has been synchronized. This search can occur in all tables of the database or only one table in the database that concentrates all the changes that have occurred; see Figure 3. The flag only changes to 0 when all modifications are finalized, i. e., when i <=n, this way if the connection is lost during this process the flag does not change, and all modification are done when the mobile unit recovers the connection.

2) *Transfer Modifications Module*

After the capture of new instructions, the modifications transfer module makes the transfer of these instructions to the server synchronization [13]. This transfer occurs according to the synchronization protocol chosen for the architecture, which, as previously mentioned, and should be a protocol that has the characteristic of interoperability and is not proprietary; see Figure 3.

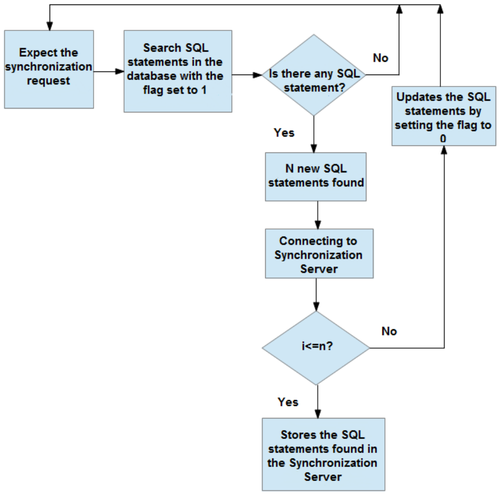

Expect the synchronization request
Search SQL statements in the database with the flag set to 1
Is there any SQL statement?
No
Yes
Updates the SQL statements by setting the flag to 0
N new SQL statements found
Connecting to Synchronization Server
i<=n?
No
Yes
Stores the SQL statements found in the Synchronization Server

Figure 3. Flow diagram of the SQL statements.Adapted from [13].

The flow diagram of these modules shows that when a connection error occurs while transferring changes, the variable n (the variable in the conditional structure of the flow diagram) will have the value of the SQL statements that were not transferred. Thus, when the internet connection is reinstated the transfer continues from where it stopped without causing data inconsistency.

V. CASE STUDY

A. *Field Data Collection Stage*

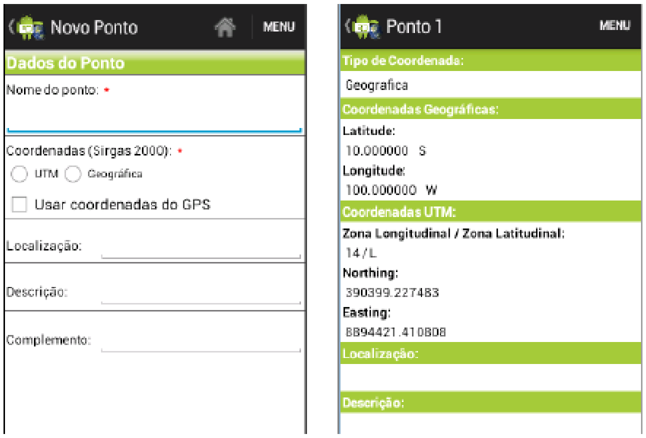
In the first stage of the architecture, for data collection in the field, a mobile application called RockDroid was developed using a methodology of data collection based on remote sensing spatial research undertaken by the Institute of Geosciences of the University of Brasilia (UNB). The purpose of the application was to facilitate the collection, storage and data synchronization using, Smartphones and Tablets, which are Mobile computational resources easily accessible to members of the collection team.

The application was developed for the Android platform, version 2.3 or later, and focused on using only libraries' open source components for easy maintenance and availability. The decision was made to use the Android platform, primarily because it is an open source operating system, which is consistent with the aim of the work, and also because the operating system is present on most Smartphones and Tablets, representing a share of 59.5% of the market in the first quarter of 2013 [5]. It also assists the developer in supporting devices of different sizes, shapes and specifications. Using some good programming practices you can develop an application that can be run on most Android devices, leaving the care system operative, resizing the visual components [7].

The RockDroid is responsible for storing information collected about geographical points, the rocks contained therein and a brief description of the specimens and structures of each rock. It provides forms for researchers to fill out with the data collected, and has screens to display information retrieved from the database, as well as displaying options to edit and delete records. The software also provides some mechanisms for validating the data entered to ensure the integrity and consistency of the database. Part of the persistence is implemented through a relational database created from SQLite [15], a small management system relational database, commonly used in embedded systems and does not require a large processing capacity [1].

Although the proposed architecture supports the collection of vector geographic data, it was used in applying the DBMS SQLite instead of SpatiaLite due to the requirements of the application, because the only vector geographic data types that the RockDroid collects are points by latitude and longitude, and so it is not necessary to use the SpatiaLite. However, the application also supports the DBMS SpatiaLite.

The user interface is responsible for displaying forms so that the user can insert or update information and display screens with data retrieved from the database. Another way to view the data was retrieved through a map on which the user can view the data geographically distributed. Another feature of the map is to display the current location of the device functionality that could be implemented through the mechanisms provided by the *Global Positioning System* (GPS) location, the cell phone networks and the Internet. Figures 4 and 5 show some images of the user interface application developed; the data are not real, but were inserted for demonstration only.



Dados do Ponto

Nome do ponto: *

Coordenadas (Sirgas 2000): *

○ UTM ○ Geográfica

□ Usar coordenadas do GPS

Localização:

Descrição:

Complemento:

Tipo de Coordenada:
Geografica

Coordenadas Geográficas:
Latitude:
10.000000 S
Longitude:
100.000000 W

Coordenadas UTM:
Zona Longitudinal / Zona Latitudinal:
14 / L
Northing:
390399.227483
Easting:
8894421.410808

Localização:

Descrição:

Figure 4. RockDroid application screens.

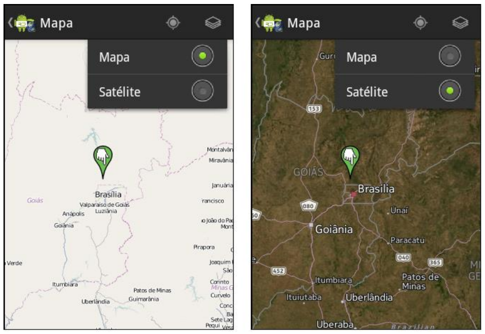


Figure 5. Data displayed on a map on the RockDroid application.

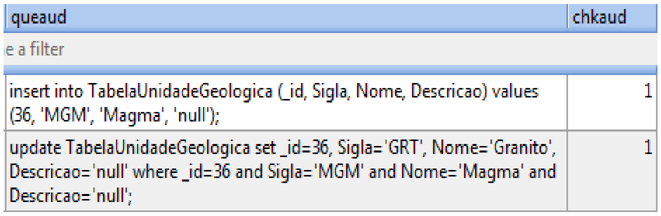
Data validation, before being entered into the database, is essential as it ensures the consistency and integrity of the information stored.

The script to create the database was embedded into the application. Thus, when installed on a mobile device, the application will create a local database. If the application is uninstalled, the database will be deleted as well. It is important to note that the database will only be created if it does not exist at the time the application starts. If there is already a database for RockDroid, it will be kept intact.

B. *Synchronization Stage*

In the synchronization stage of the architecture, the capture module modifications were developed based on the creation of triggers that are fired when there is an insert operation, update, or delete the database. These triggers aim to store all the changes in the database in one table, facilitating the search for such modifications. All changes are stored in the audit table that has the following columns: date and time of the operation (timestamp of the operation, dthaud), the database where the action occurred (banaud), the table where the action occurred (tabaud), the type of operation that occurred (tacaud), the query executed during the action (queaud), and a flag that indicates whether the record has been synchronized or not (chkaud), as explained in the previous section. Figure 6 shows a section of the table with some audit records stored. Initially this flag is set to 1 for all changes because none of them had been synchronized with the synchronization server yet.

The SyncML protocol [12] was chosen for use in the proposed architecture because this protocol meets the prerequisites that the proposed architecture requires, i.e., promotes interoperability of data synchronization and is a non proprietary protocol. It is a standard protocol for data synchronization, regardless of platform, device and network. It is based on Extended Markup Language (XML) technology and maintained by Open Mobile Alliance (OMA), an alliance of several companies to create a common protocol for data synchronization [9][10].



queaud	chkaud
e a filter	
insert into TabelaUnidadeGeologica (_id, Sigla, Nome, Descricao) values (36, 'MGM', 'Magma', 'null');	1
update TabelaUnidadeGeologica set _id=36, Sigla='GRT', Nome='Granito', Descricao='null' where _id=36 and Sigla='MGM' and Nome='Magma' and Descricao='null';	1

Figure 6. A piece of the table with two audit records stored.

In the proposed synchronization, we used the so-called one-way synchronization from the client to the server, where the changes in the database of mobile researchers (SyncML clients) are synchronized with the server synchronization (SyncML) [9][10]. This type of synchronization was chosen because in a data collection it is necessary to have all the data collected only in the server, due to collection devices having limited computational resources.

The transfer module of modifications has been developed in Java using the library server Sync4j Funambol Data Sync Server (Funambol). The Funambol is a synchronization server that implements the SyncML protocol and was used to implement the synchronization server architecture [2]. This module is nothing more than a SyncML client to query the audit table, capturing the changes in the database, and transfers the changes occurring to the synchronization server. After synchronization, the flags of the changes are synchronized to 0 unset.

C. *Storage Stage*

The last stage, i.e., the storage stage, was developed through replication (backup) server data synchronization (Funambol) to the server database, which was developed based on the PostgreSQL DBMS with spatial extension PostGIS.

D. *Results*

Three tests were performed on a heterogeneous computing environment. The three tests performed were:

- Check if synchronization occurs properly and in a timely fashion;
- Check if, when a disconnection of the internet in the middle of a sync occurs, the sync continues where it left off or resynchronizes after the reestablishment of the connection ;
- Check if the synchronization of multiple mobile devices simultaneously occurs correctly (we used five mobile devices for testing).

Figure 7 shows the computing environment in which the proposed architecture was tested.

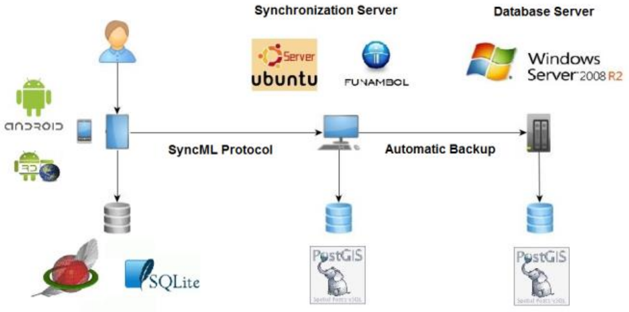


Figure 7. The heterogeneous computing environment in which the proposed architecture was tested.

Phase of data collection in the field:

- Device used for collecting data: tablet and smartphone;
- Operating System: Android;
- Computer system: geographic information system (RockDroid);
- Supported data types: conventional and geographic data;
- System manager database: SQlite and SpatiaLite.

Synchronization Phase:

- Type of computer used as synchronization server: desktop;
- Operating System: Linux Ubuntu Server 12;
- System manager database: PostGIS;
- Supported data types: conventional and geographic data;
- Protocol synchronization: SyncML (Funambol).

Storage phase:

- Type of computer used as server database: Minicomputer;
- Operating System: Windows Server 8 R2;
- System manager database: PostGIS;
- Types of data stored in the server database: conventional and geographic data.

The synchronization was tested in different types of networks (wireless, wired) and in different states of the network (Internet with volume data traffic 600 megabits per second, 100 megabits per second and 10 megabits per second), and in every case both the collection and the synchronization were successful and presented acceptable times, respectively 1 to 3, 4 to 8 and 8 to 10 seconds; the time synchronization varied according to the state of the network.

Tests were conducted off the Internet in the middle of a synchronization and after the reestablishment of the connection, synchronization restarted again. Both the capture module modifications and the transfer module modifications followed the diagram shown in the previous section, so there was fault tolerance.

The last test was to synchronize multiple mobile devices simultaneously. This test was intended to simulate the reality of the collection, because normally all researchers, after collection, synchronize their data simultaneously to the server. Five mobile devices were synchronized at the same time and no error occurred when synchronizing.

VI. Conclusion and Future Work

The existing data collection architectures satisfy a specific cause, using specific technologies that are limited regarding the type of data, type of networks, synchronization type, device type, and so on. For each different data collection context, an architecture is proposed for data collection with different budget limitations, computational environments and available technology within the company or organization.

The proposed architecture for this work was an architecture of data collection for field research in heterogeneous computational environments that supports vector geographic data where information stored on each mobile device are replicated and correctly integrated into a central database. To reach this goal, the architecture had to obey two aspects: interoperability and flexibility. This architecture complies with the aspect of interoperability because it uses different mobile devices, operational systems and wireless network properties. It also complies with the aspect of flexibility, by replicating data amount where there is an intermediate server and a free synchronization protocol, which supports replication to many databases. In addition, the architecture uses some safety mechanisms such as fault tolerance, full replication data and authentication.

With this architecture, it is no longer necessary to plan an architecture whenever you make a data collection in the field. Researchers may simply use the proposed architecture instead, which in addition to functioning in heterogeneous environments and support vector geographic data, is low cost and high quality.

Further testing of these applications is will be carried out in the field. These tests will be conducted with the support of the Institute of Geosciences, University of Brasília and concerns the collection of geological data in several cities of Brazil.

References

[1] About SQLite, http://www.sqlite.org/about.html. [retrieved: Feb., 2014]

[2] A. L. B. Alonso, C. Oliveira, L. Fedalto, F. Vilas Boas, T. L. G. Assis, and C. S. Hara, "A synchronization experience of relational databases using SyncML.", In: . Proc.: Escola Regional de Banco de Dados (ERBD), Brazil, Apr. 2010, pp. 1-4.

[3] AuditMatic, http://www.auditmatic.com. [retrieved: Feb., 2014]

[4] B. R. Badrinath, and S. H. Phatak, "Bounded locking for optimistic concurrency control," Department of Computer Science, University Rutgers, New Jersey, EUA, 1995.

[5] Canalys, http://www.canalys.com/newsroom/smart-mobile-device-shipments-exceed-300-million-q1-2013. [retrieved: Feb., 2014].

[6] D. L. Costa and F. Franquini, "Synchronization protocols in wireless environment," Department of Computer Science of Federal University of Santa Catarina, Master These, Brazil, 2004

[7] A. Developers, http://developer.android.com /guide/practices/screens_support.html.[retrieved: feb., 2014]

[8] R. Elmasri, and S. B. Navathe "Fundamental of Database System," Addison Wesley, USA, 2011.

[9] Ericsson, IBM, Lotus, Matsushita, Communications Industrial Co. Ltd., Motorola, Nokia, Openwave, Palm, Psion, Starfish Software, Symbian, and others "Building an industry-wide mobile data synchronization protocol," SyncML White Paper, 2000.

[10] Ericsson, IBM, Lotus, Matsushita, Communications Industrial Co. Ltd., Motorola, Nokia, Openwave, Palm, Psion, Starfish Software, Symbian, and others "SyncML sync protocol," SyncML White Paper, 2002.

[11] S. Gejibo, F. Mancini, K. A. Mughal, R. A. B. Valvik, and J. Klungsøyr "Secure data storage for mobile data collection systems,". Proc: IEEE HealthCom conference, 2012, pp. 498-501.

[12] U. Hansmann, R. Mettala, A. Purakayastha, P. Thompson, and P. Kahn "SyncML: synchronizing and managing your mobile data," Prentice Hall, 2002.

[13] M. I. Hossain, and M. M. Ali "SQL query based data synchronization in heterogeneous database environment," International Conference on Computer Communication and Informatics, Coimbatore, India, Jan. 2012, pp. 1-5.DOI: 10.1109/ICCCI.2012.6158818

[14] G. C. Ito, M. Ferreira and N. Sant'Ana, "Mobile Computing: characteristics about data management ", Instituto Nacional de Pesquisas espaciais – INPE, 2003.

[15] R. Lecheta "Aprenda a criar aplicações para dispositivos móveis com o Android SDK," NOVATEC, São Paulo, SP, Brasil, 2009.

[16] J. Magalhães, M. Holanda, and R. Chaim "Architecture for data collection in mobile computing with intermittent internet access," 6ª Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI 11), Chaves, Portugal, AISTI Press,2011, pp. 1-6.

[17] E. C. Manganelli, and J. Romani "Data synchronization protocols in wireless environments: a case study," Department of Computer Science, Master These, Federal University of Santa Catarina, SC, Brazil, 2004.

[18] Nokia Data Gathering, http://www.nokia.com/corporateresponsibility/society/nokia-data gathering/english. [retrieved: Feb., 2014]

[19] M. Rennhackkamp "Mobile Databases Tetherless Computational Liberates End Users But Complicates the Enterprise," DBMS online, 1997.

[20] A. J. S. Silva, A. S. de Andrade Júnior, and F. R. Marin "Data Collection and dissemination Architecture of climate data in the state of Piauí," Revista de Tecnologia de Fortaleza, vol. 29, Brazil, 2008.

[21] M. Ji, Y. Sun, F. Jin, T. Jiang, J. Wang, and X. Yao "Research and Development of Field Data Collecting Synchronously System of Mining Area," IEEE Internacional Geoscience and Remote Sensing Symposium, Hawaii, USA, IEEE Press, 2010. pp. 3948 – 3951. DOI: 10.1109/IGARSS.2010.5650825.

Integrative Analysis of Land-use and Road Network Structure

Hyeyoung Kim, Chulmin Jun
Dept. of Geoinformatics, Univ. of Seoul
Seoul, Korea
e-mail: {lucykhy, cmjun}@uos.ac.kr

Abstract— **In urban spaces, road and land use interact with each other. To understand urban spaces and make appropriate plans, integrative analyses considering these two simultaneously are required. So far, advancement in urbanization has led to a higher building density, rather than to a sprawl of urban areas. Therefore, the purpose of this study is to conduct integrative analysis and evaluation of road and land use considering the characteristics of modern cities. Based on road as an appropriate accessibility variable, modified importance performance analysis was conducted with development density and the results were categorized into four areas. To apply appropriate accessibility variable, space syntax theory considering the structure of road network was introduced for road. For land use, to consider both horizontal and vertical development densities of residential and commercial buildings were used. The proposed method was applied to Gangnam-gu, a central business district area in Seoul, and results were analyzed and visualized using geographic information system.**

Keywords-road network; space syntax; development density; IPA; integrative analysis.

I. INTRODUCTION

Road is closely related to land use and the structure of road composition significantly affects the structure of land development. For sustainable urban development, integrative analysis considering road and land use simultaneously is required.

Modern cities are in the structure of compact cities in which high density developments are proceed rather than urban space expand outward. For the proper analysis of urban spaces, not only their horizontal aspects but also their vertical aspects should be examined. In relation to this, previous studies mainly used data, such as populations and employees for analyzing urban density [1][2]. However, such data are usually summed up by administrative districts and thus, they are not proper for detailed analysis related road structure at finer level. As accessibility related data, Euclidean distances between two areas or facilities have been widely used [3][4]. However, using Euclidean distances that measure the direct distances between two points have limitations in that standards for to and from points do not exist and they do not reflect the form of road network. Recently, studies have been conducted in relation to space syntax in which the structure of road network based on visibility are considered. These studies showed that attributes calculated by space syntax are closely related to city components [5][6][7][8]. Although there was a study on transformation of urban patterns through analysis of urbanization rate with space syntax [9], studies that conducted integrative analysis and utilization of the attributes in combination with land use are insufficient.

The purpose of this study is to conduct integrative analysis and evaluation of road and land uses considering the characteristics of modern cities where road and land-use developments interact with each other. First, for road network, global integration was used among attributes calculated by the space syntax theory considering the structure of road network. For land use, to consider both horizontal and vertical development densities, the building plan areas and gross floor areas of residential and commercial buildings were used. In addition, to apply appropriate accessibility variables that would become criteria for analysis, the explanatory power of three variables, namely, Euclidean distance, global integration and length-reflected global integration were compared. The explanatory power of these variables was expressed using development density and land price. Finally, a modified importance performance analysis model based on road as an accessibility variable was conducted with development density of residential and commercial land uses and the results were visualized. This study was tested on 22 administrative districts in Gangnam-gu of Seoul City, a planned central business district area developed in the 1970s.

The paper begins with introduction of research. In Section 2, methodology of space syntax theory and modified IPA and data construction is described. The explanatory power of accessibility variables were compared in Section 3. Then results of IPA were expressed in Section 4. Concluding remarks are given in Section 5.

II. METHODS AND DATA CONSTRUCTION

A. *Methodology Space Syntax theory*

Space syntax is a method that analyzes relative accessibility quantitatively based on visibility of roads recognized by humans [10][11][12]. Space syntax uses ‘the depth’ of spaces instead of Euclidean distances in order to compute the attribute values of the model. The depth means the minimum number of connecting lines that should be gone through when moving from a certain space to another. The depth between adjacent spaces is 1 and it increases as the levels where a space to pass to another increases. In space syntax, axial maps should first be prepared with the minimum number of lines that connect the longest possible straight lines with each other as shown in Figure 1. The attribute values are calculated and assigned in axial lines instead of intersecting points as with transportation network.

In this study, global integration was used among the attributes of space syntax. This attribute is calculated based on roads that connect all axial lines in the scope of the subject of analysis assuming that the axial lines are starting points and end points and the depths. A large value of global integration in a certain space means that the numbers of axial lines that are gone through to move all other spaces are relatively small. That is, a large value of global integration in a certain space means that the space is the center of all spaces, accessibility to all other spaces is high, and movements to all other spaces are easy.

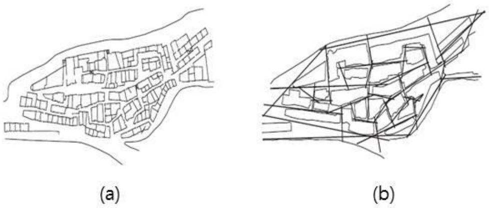


Figure 1. Preparation of axial map: (a) Subject of analysis (b) Axial map

B. *Modified IPA*

Importance-Performance Analysis (IPA) was developed by Martilla & Jame to establish efficient strategies to invest limited resources in the area of marketing [13]. In this analysis, importance and performance are expressed with two dimensional graphs having x and y axes and each of the graphs are divided into four areas based on the average value of each axis. That is, this is an evaluation method for comparing and analyzing the relative importance and performance of goods or services simultaneously.

In this study, a modified IPA was developed and conducted for integrative analysis and evaluation of roads and land uses in urban spaces. The X and Y axes were made to indicate road network and development density respectively. More rightward points on the X axis indicate higher accessibility and higher points on the Y indicate higher development density. The Cartesian space in Figure 2 was made using average values by administrative districts on individual axes and development density was defined into four areas as follows based on road network.

- Area 1 (Density-Road Balanced):
  Both accessibility and development density are high.
- Area 2 (Density > Road):
  Development density is higher compared to accessibility.
- Area 3 (Low Density-Road):
  Both accessibility and development density are low.
- Area 4 (Density < Road):
  Development density is lower compared to accessibility.

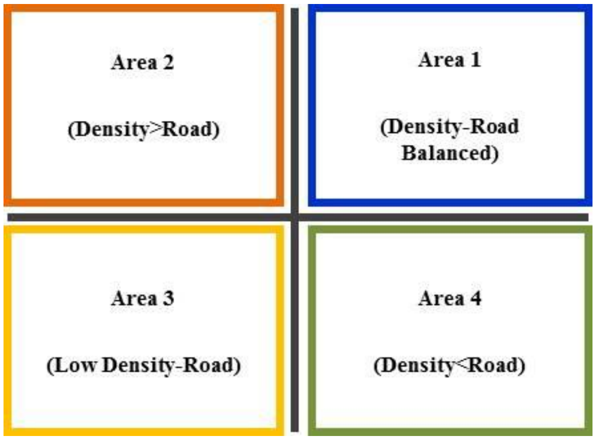


Figure 2. Definition of modified IPA area

C. *Study area*

In this study, Gangnam-gu was selected as a study area, which is approximately 10 km away from the Han-river to the southeast and plays the role of a sub-center of Seoul. Gangnam-gu is a planned area for which land compartmentalization rearrangement projects were implemented in the 1970s in which wide orthogonal main roads were developed except for greenbelts on the southern part. Currently, Gangnam-gu consists of 22 administrative districts and its area is 39.55 km² that corresponds to 6.53% of Seoul. Of the area, 56.28% is residential areas, 6.14% is commercial areas and greenbelt areas. Therefore, land uses are relatively in harmony with balance. The residential areas are mostly large apartment complexes or multi-unit dwellings and have been evenly developed throughout the entire area and commercial areas include few industrial regions. Figure 3 shows road network and development density of Gangnam-gu.

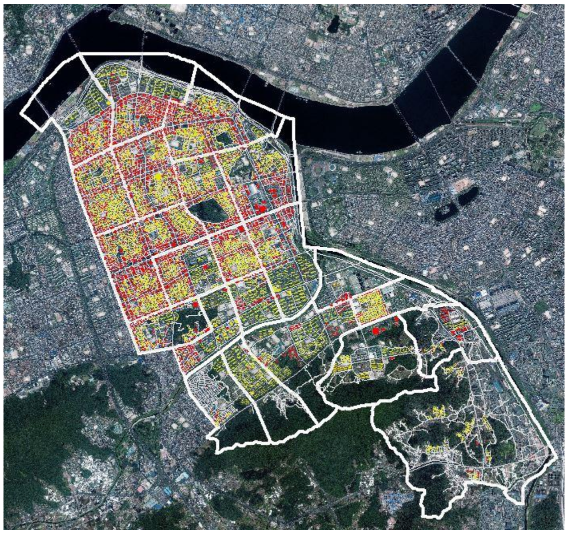


Figure 3. Gangnam-gu road network and development density

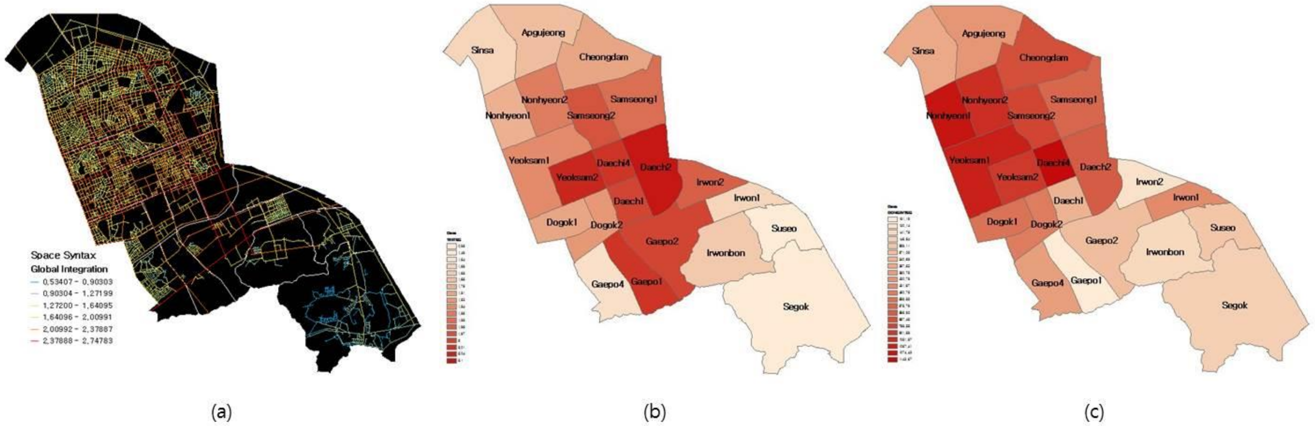


Figure 4. Construction and processing of global integration: (a) Gangnam-gu axial map (b) Global integration (c) Length-reflected global integration

D. Generation and storage of the simulation results

In this study, to analyze the density and structures of spaces, the 2009 Seoul KOTI spatial data and the 2006 publicly notified individual land price were utilized. The average area of 22 administrative districts in Gangnam-gu is 1.8 km². Because the sizes of space units vary from 0.73 to 6.36 km², the constructed data were processed in proportion to areas.

1) Accessibility variable: As accessibility variables, Euclidean distance, global integration, and length-reflected global integration were constructed.

a) Euclidean distance: This is the linear distances from the center of each parcel to the nearest road link. The averages of the distances were calculated as follows.

$$E_A = \sum_{i=0}^{n} PD_i \tag{1}$$

where E_A is the average of distances of parcels to roads of area *A*, PD_i is distance of parcel *i* to the nearest road, and *n* is the number of parcels in area *A*.

b) Global Iintegration: To construct global integration, as mentioned earlier, the axial maps of Gangnam-gu were built using Axwoman as shown in Figure 4-(a). The global integration is a value obtained by (2).

$$G_A = \frac{\sum_{i=0}^{k} I_i}{k} \tag{2}$$

where G_A is the average global integration of area *A*, I_i is the global integration of road *i*, and *k* is total number of roads (axial lines). Figure 4-(b) shows the calculated values that are between the highest at 2.1 and the lowest at 0.98. The values above 1 mean the strong 'integration', while the values between 0.4 and 0.6 show somewhat 'segregation'.

c) Length-reflected Global Integration: The space syntax stemmed from the study of architecture, and the related studies have shown that the element 'depth' had higher explanatory power than 'Euclidean distance' [14]. Spaces syntax reflects the structure of road network in the computing but it excluds actual capacity (*e.g.*, length) of roads. However, since the centers of urban spaces have higher road capacity than the outskirts, weighted values should be given in accordance with areas. Therefore, as the third variable, the global integration that reflected road lengths was calculated as shown in (3).

$$GL_A = \sum_{i=0}^{k} L_i I_i \tag{3}$$

where GL_A is the global integration with road length of area *A*, and L_i is length of road (axial line) *i*, I_i is global integration of road *i*, and *k* is total number of roads. Figure 4-(c) shows the calculated values that are between the highest at 1143.67 and the lowest at 121.12. The average value was shown to be 535.51.

2) Development density and land price: Development density values were obtained for residential and commercial uses separately. Plan areas and gross floor areas of buildings were processed to analyze horizontal and vertical densities. Plan areas were obtained from building data and gross floor areas were calculated by multiplying plan areas by the numbers of building floors. The average plan areas of residential and commercial buildings are 754.75 and 631.20 respectively and the average gross floor areas of residential and commercial buildings are 5484.61 and 4204.41. Figure 5 shows the values of residential and commercial buildings devlopment density by administrative districts. The effects of land price on accessibility variables explanatory power and the development of residential and commercial areas were examined. Land prices were calculated as averages weighted by the ratio of individual parcels.

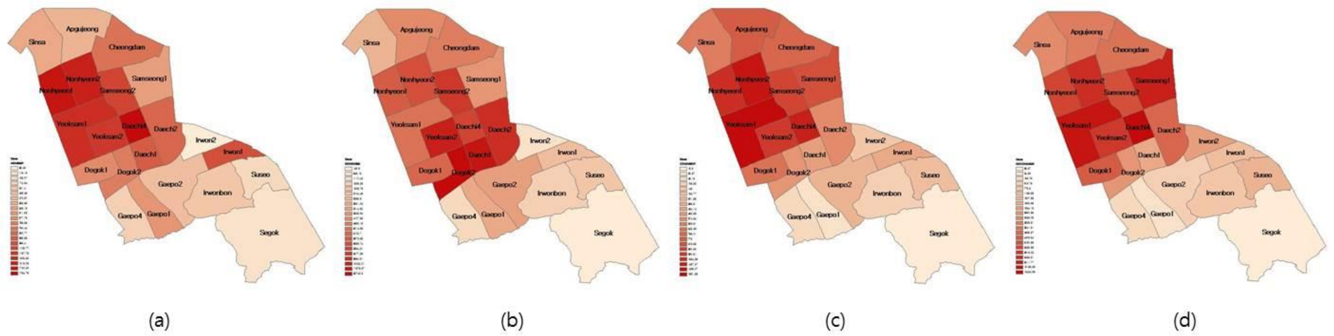


Figure 5. Construction and processing of development density: (a) Residential buildings plan areas (b) Residential buildings gross floor areas (c) Commercial buildings plan areas (d) Commercial buildings gross floor areas

III. Explanatory Power of Accessibility Variables

To apply appropriate accessibility variables, the correlations of Euclidean distance, global integration and length-reflected global integration were analyzed each with development density and land price as shown in Table I. The Euclidean distance showed significant correlation only with commercial density and showed negative (-) relationships. This means that the shorter the distances for access of areas are, the higher the development density of the areas is. The global integration showed relatively high correlation with land prices and residential density and showed higher explanatory power for land prices. The length-reflected global integration had high correlation and explanatory power in most cases. In particular, its correlation with commercial density showed high explanatory power while the correlation with residential density showed relatively low explanatory power for gross floor areas compared to plan areas due to high residential density where many apartments are located. In general, commercial uses are greatly affected by accessibility and commercial zones are formed along the roads. The analysis also showed that commercial development is more affected by the lengths of roads than other factors.

TABLE I. Explanatory Power of Accessibility Variables

		Euclidean distance	**Global integration**	**Length-reflected global integration**
Land price		-0.361 (<0.99)	0.686 (<0.01)	0.636 (<0.01)
Plan area	**Residential**	-0.392 (<0.071)	0.460 (<0.05)	0.910 (<0.01)
	Commercial	-0.551 (<0.01)	0.361 (<0.09)	0.946 (<0.01)
Gross floor area	**Residential**	-0.166 (<0.46)	0.432 (<0.05)	0.341 (<0.12)
	Commercial	-0.452 (<0.05)	0.372 (<0.08)	0.874 (<0.01)

Therefore, the length-reflected global integration was selected as an appropriate accessibility variable to be applied to the study.

IV. Results of IPA

As mentioned earlier, IPA was conducted for integrative analysis of road network and development density and the results for individual areas were visually expressed. The x-axis was defined as length-reflected global integration in road network and the y-axis as development density or land price.

A. *Road network-residential density*

The x-axis was defined as road network and the y-axis was residential building plan areas and gross floor areas relatively as shown in Figure 6 (a) and (b). Area 1 is blue areas where both accessibility and residential development density are high. Area 2 represented with red is the areas that show residential development density is higher compared to accessibility. Area 3, yellow color areas, shows both low accessibility and low residential density. Area 3 includes most districts in the southern part of Gangnam-gu. The reason that these areas show relatively low development density is because large greenbelts are located in the southern part. Area 4, in green color, show that residential development density is lower compared to accessibility. Although this area has low residential density compared to its accessibility, its commercial density seems to be high. On comparison of residential buildings, it can be seen that Cheongdam, Yeoksam 1, and Ilwon 1 districts moved to areas 3 and 4. This means that the gross floor areas of residential buildings are relatively small in these areas.

B. *Road network-commercial density*

The x-axis was defined as road network and the y-axis was commercial buildings plan areas and gross floor areas, as shown in Figure 6 (c) and (d). Area 1 is blue areas where both accessibility and commercial development density are high. Most districts in the central part of Gangnam-gu are included in this area. Area 2 is represented with red is the areas that show commercial development density is higher compared to accessibility. Area 3 includes the yellow areas where both accessibility and commercial density are low.

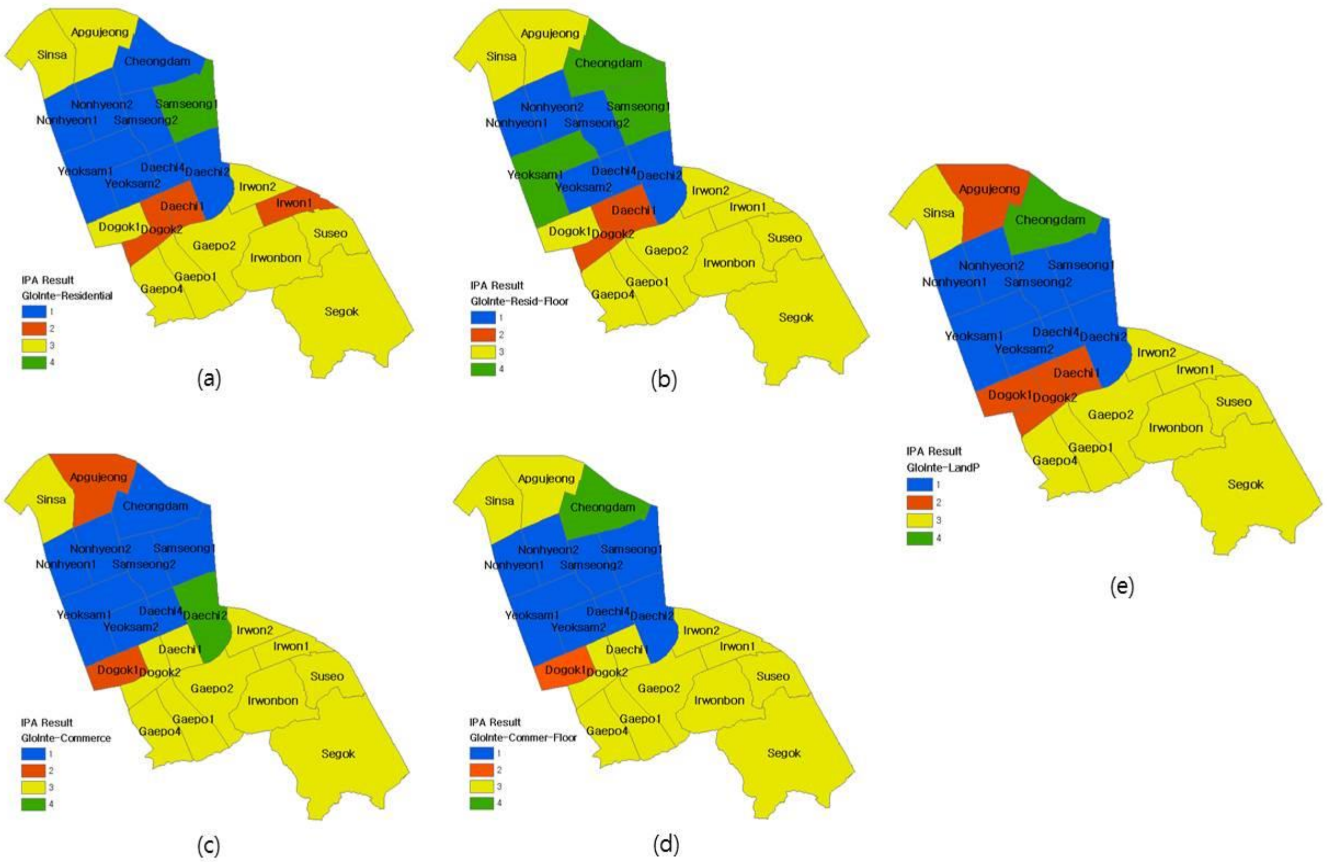


Figure 6. Results of modified IPA: (a) Residential buildings plan areas (b) Residential buildings gross floor areas (c) Commercial buildings plan areas (d) commercial buildings gross floor areas (e) land price

Since the ratio of greenbelts was high, commercial density in this area was shown to be similar to residential density. Area 4, with green color, show that commercial development density is lower compared to accessibility. Although Daechi 2 district is included in this area, the gross floor area of commercial buildings was shown to be relatively large. On the other hand, Cheongdam and Apgujeong districts were shown to have relatively small gross floor areas of commercial buildings.

C. Road network-land price

The x-axis was defined as road network and the y-axis was land price as shown in Figure 6-(e). Based on road network, how the effects of land price are applied to development density can be examined. Most areas are included in the same areas based on both land price and development density. To review those districts that were included in different areas by residential and commercial density, land price in Daechi 1, Dogok 2, and Daechi 2 districts were affected by residential density, those in Dogok 1, Samsung 1, and Apgujeong districts were affected by commercial density, and those in Cheongdam district were affected by the gross floor areas of residential and commercial buildings.

V. CONCLUDING REMARKS

In this study, integrative analysis of roads and land use was conducted with 22 administrative districts in Gangnam-gu, is a planned CBD. In order to consider the characteristics of the city, spatial data such as road structures and development density were used. Space syntax theory was employed for computing road structure and global integration was used among its attributes. For land use, to analyze both horizontal and vertical densities, the building plan areas and gross floor areas of residential and commercial buildings were used.

The levels of explanatory power of Euclidean distance, the global integration, and the length-reflected global integration were compared with each other. The correlations between these accessibility variables and development density and land price were analyzed. According to the results of the analysis, the length-reflected global integration showed low explanatory power for residential buildings gross floor areas but showed high explanatory power for residential buildings plan areas. In particular, it showed high explanatory power for commercial density which is closely related to accessibility. Therefore, it was selected as an appropriate accessibility variable.

A modified IPA model that use the length-reflected global integration and development density was conducted and the results were visualized using GIS. Gangnam-gu was classified into four areas based on average values on individual axes and the areas were relatively analyzed. These four areas were defined as ‘density-road balanced area’, ‘density > road area’, ‘low density-road area’, and ‘density < road area’ respectively; also, how the effects of development density acted on land price were identified. It is viewed that, with more refinement, the method suggested here can be used in analyzing urban development in microscopic level considering road structure.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government. (No. 2011-0009601)

REFERENCES

[1] D. I. Lim, “Analysis on the Changes of Urban Structure by New Town Development,” The Journal of the Korea Contents Association, vol. 8, no. 10, 2008, pp. 317-327.

[2] G. W. Lee and S. Y. Kim, “A Study on the Changes of Suwon city’s Urban Spatial Structure,” Asian pacific Planning Review, vol. 45, no. 1, 2010, pp. 7-20.

[3] S. B. Lee, “The correlation analysis between land prices and accessibility and land use zones in the north Cheonan city,” Korea East West Economic Research, vol. 17, no. 2, 2006, pp. 59-77.

[4] W. K. Min, “A study of a land special quality effect on Posted land Price,” Residention Environment Institute of Korea, vol. 4, no. 1, 2006, pp. 99-113.

[5] B. Hillier, R. Burdett, J. Peponis, and A. Penn, “Creating life: Or, does architecture determine anything?,” Architecture and Behaviour, vol. 3, no.3, 1987, pp. 566-250.

[6] J. Peponis, C. Zimring, and Y. K. Choi, “Finding the building in wayfinding,” Environment and Behavior, vol. 25, no. 5, 1990, pp. 555-590.

[7] H. L. Kim and D. W. Sohn, “An analysis of the relationship between land use density of office buildings and urban street configuration,” Cities, vol. 19, no. 6, 2002, pp. 409-418.

[8] Y. W. Kim, “A study on the relationshop between properties of spatial configuration and patterns of space use using space syntax,” Asian pacific Planning Review, vol. 38, no. 4, 2003, pp. 7-17.

[9] H. Y. Kim, Y. J. Joo, and C. M. Jun, “A study on transformation of urban layout patterns through analysis of spatial relationships with urban street configurations,” International Journal of Urban Sciences, vol. 15, no. 1, 2011, pp. 25-34.

[10] B. Hillier and J. Hanson, “The Social Logic of Space,” Cambridge University Press, 1984.

[11] B. Hillier, “Space is the Machine,” Cambrige University Press, 1996.

[12] B. Hillier, “Space is the Machine,” University of Cambridge Press, 2007.

[13] J. A. Martilla and J. C. James, “Importance-Performance Analysis,” Journal of Marketing, vol. 41, no. 1, 1997, pp. 77-79.

[14] Y. W. Yun, Y. Y. Kim, and Y. K. Park, “A study on Influence by Depth and Distance in Spatial Cognition,” Journal of Architectural Institute of Korea, vol. 23, no. 1, 2003, pp. 171-174.

Analysis of Clustering and Unsupervised Learning of Geospatial Demographic Data

Mikhail Kanevski and Jean Golay
Institute of Earth Surface Dynamics
Faculty of Geosciences and Environment
University of Lausanne, Switzerland
Mikhail.Kanevski@unil.ch, Jean.Golay@unil.ch

***Abstract*—The research discusses the methodological framework of an application of a newly developed spatial clustering algorithm – the functional multipoint Morisita index (fm-Morisita) - and an unsupervised learning algorithm – self-organizing Kohonen maps (SOM), for the comprehensive exploratory analysis and quantification of patterns in high resolution geospatial demographic data in Switzerland. fm-Morisita is used to analyse the complex clustering of the spatial distribution of the population. The SOM are used to reveal regional patterns of similarity using detailed information about ageing groups.**

Keywords - geodemography; unsupervised learning; spatial data clustering.

I. Introduction

The basic motivation for this research is to explore the structure and evolution of demographic and other geospatial data of the “SuisseMetropole” (SMP) using cutting-edge methods and algorithms from machine learning (mainly unsupervised learning and dimensionality reduction tools [1,2]), geocomputation (city clustering algorithms [3]) and spatial clustering using topological, statistical and fractal measures [4]. In this work, urban zones of Switzerland are considered as a “SuisseMetropole”, i.e. a complex multivariate, multiscale (from hectometric intra-city to inter-city level), and high dimensional hierarchically organized system. One of the main intentions is to rely on real geodemographic data as much as possible, following the principle “let the data speak for themselves” in making as few prior hypotheses and assumptions as possible by following an unsupervised learning approach. Geocomputational approaches will help to carry out analysis at multiple selected scales and to better understand the structural and functional optimality of the “SuisseMetropole”. A subsequent evaluation of the relevance of the results for the fields of urban, economic and social geography will be carried out.

Fundamentally, the first problem is to quantify and to understand the clustering of populated areas using a recently developed method – the functional multipoint Morisita index (fm-Morisita). The second one deals with the analyses of data in a high dimensional space (dimensionality >10-100) in order to understand the geodemographic patterns of “SuisseMetropole”. Below, the data and the main methods proposed and how they will be applied are shortly discussed.

II. Data

The Swiss Federal Statistical Office (OFS) and the Swiss Federal Office of Topography (Swisstopo) have provided the main data required by the research plan. OFS data are organized on a regular grid of high resolution (100m x 100m) and they can be subdivided into different sections. For the present research, the following data sets were mainly used: population census of years 1990, 2000 and 2010. They contain variables about the demographic and socio-economic characteristics of the population: resident population organized by language, place of residence, gender, age (19 classes), employment; information about education, socio-economic status, mean of transportation and households composition. The global distribution of the population in 2010 is given in Fig.1. The simulated data with known clustering structures (i.e. complete spatial randomness) were generated within the validity domains and the results were compared with the original geodemographic data (see Fig.2).

III. Methods

Complex and challenging data demand the application of advanced data modelling tools, including machine learning algorithms.

In the present research, the two main methods applied for the study of Swiss complex geodemographic data are based on 1) machine learning – self-organizing maps (unsupervised learning) and 2) geospatial data clustering algorithms –the Morisita index and fractal measures. [1,2,4,5,6]. The basic idea is to find and to model geospatial patterns (structures in space) in the high resolution demographic data. The first approach - SOM is used to reveal patterns in the 19th dimensional space of the Swiss demographic data composed of 19 classes of age groups (0-5, 5-10,…,>90 years) and to visualize them using Geographic Information System in a two dimensional space. SOM is a well know exploratory analysis and visualization tool making a topology preserving mapping (projection) from high dimensional space to two dimensional space. An important advantage of SOM is that it is an unsupervised algorithm and does not need a priori knowledge about the number of similarity clusters in data. Here, the geographical information (i.e.the coordinates) were left aside and only the demographic data were used.

The second objective of the research takes into account the distribution of the population in the two dimensional geographical space using fm-Morisita and fractal measures of clustering in order to understand the spatial distribution of the populated areas [5]. In [5], fm-Morisita method was first time introduced for spatial data and it's the relationships of the multipoint Morisita index (on which fm-Morisita is based) with multifractality was demonstrated.

The multipoint Morisita index is calculated as follows: the region of interest in covered by non-overlapping cells of size l and the probability to find m points in a cell is calculated. Cell size and m are the free parameters. When taking into account the density of the population (not only the distribution in space) functional measures (depending on local density) based on the multipoint Morisita index can be implemented. The details can be found in [4,5]. In order to demonstrate that observed patterns are really structured, a comparison with point distributions generated generated by shuffling the original data was carried out.

In the present research, both methods were applied for the first time for high resolution demographic geospatial data.

IV. Results and Conclusions

The preliminary results of the first analysis (SOM) are quite interesting and demonstrate that several demographic spatially structured patterns can be observed in Switzerland: dense urban zones, interurban regions and country patterns. More detailed analyses will also take into account also the level of education and other relevant socio-economic parameters.

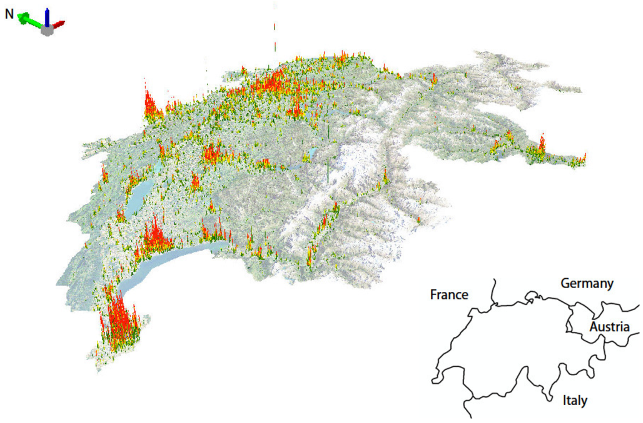


Figure 1: Distribution of the Swiss population in Year 2010.

The application of fm-Morisita has confirmed that the clustering of population in Switzerland (SuisseMetropole) is of multifractal nature.

In future research, the methods will be applied to several temporal datasets, which will help to understand the evolution of the detected geodemographic patterns in time. Moreover, multivariate measures – joint multifractals and multivariate fm-Morisita-will be analysed to relate demographic and socio-economic data.

Acknowledgment

The preliminary research was partly supported by the Faculty of Geosciences and Environment, University of Lausanne within the framework of the FINV "SuisseMetropole" project.

References

[1] T. Kohonen. "Self-Organizing Maps". Springer, 2001.

[2] M. Kanevski, A. Pozdnoukhov and V. Timonin. "Machine Learning for Spatial Environmental Data". EPFL Press, 2009.

[3] H. Rozenfeld, Rybski D., Andrade J, Batty M., H. Eugene Stanley H. E., and Makse H. "Laws of population growth". PNAS, vol. 105, No. 48, 2008.

[4] M. Kanevski (Editor). "Advanced Mapping of Environmnetal Data". iSTE and Wiley, 2008.

[5] J. Golay, M. Kanevski., C. D. Vega Orozco, and M. Leuenberger. "The Multipoint Morisita Index for the Analysis of Spatial Patterns". Available on arXiv (arxiv.org):1307.3756, 2013 [accessed March 2014].

[6] C. D. Vega Orozco, J. Golay, and M. Kanevski "Multifractal portrayal of the Swiss population" Available on arXiv (arxiv.org): 1308.4038, 2013 [accessed March 2014].

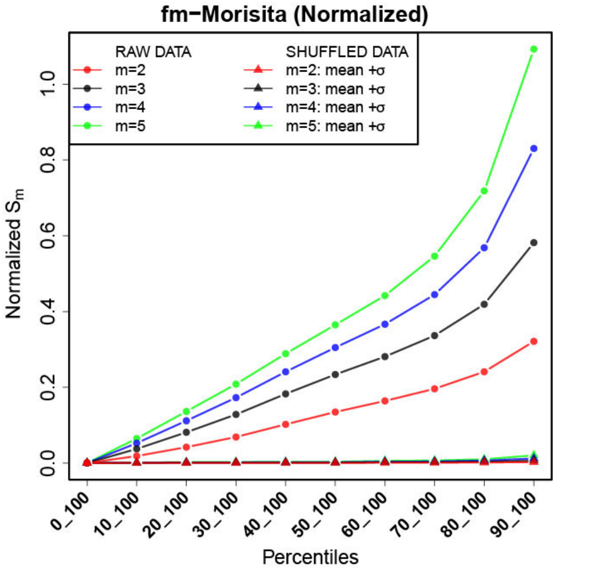


Figure 2: Application of fm-Morisita [3] to the distribution of the Swiss population. The X-axis corresponds to the percentage of the population above the mentioned deciles (e.g., "20_100" means that all the hectares associated with a value lower than the second decile are rejected) and the Y-axis is a normalized index derived from the m-Morisita index. The results for shuffled data are given as well. They partially overlap.

Air Pollution Dispersion Mapping by Remote Sensing: Case Study from the Federal District

Erick Frederico Kill Aguiar, Henrique Llacer Roig

Institute of Geosciences, University of Brasília - UnB

Brasília, Brazil

E-mails: {erickkill@aluno.unb.br, roig@unb.br}

***Abstract* - The city densification and, consequently, the potential polluting activities have increased significantly in the last decades, compromising the air quality. Thus, it is necessary to know this scenery to model the pollution behavior and its respective pattern in each region of the Federal District. So, the main goal of this study was to analyze the Aerosol product available by National Aeronautics and Space Administration (NASA) at the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor (TERRA). With these preexisting data, it was possible to analyze the aerosol (AOD) distribution from 2010 and the optical depth. These data were integrated to the processed values of Total Particles in Suspension (PTS) of the eight air quality monitoring stations variably distributed in the Federal District, with two of them (714 south and HUB) kept and managed by the Interdisciplinary Center of Transport Studies (Centro Interdisciplinar de Estudos em Transportes - Ceftru/UnB), correlating the month of October 2010. The study showed that the region of Brasília downtown and the surroundings of the Fercal region, overall, were the ones that kept most of the Aerosol Optical Depth (AOD) and the highest values of the Angström Exponent (AE), which can be attributed to the types of activities performed in these regions.**

Keywords: Aerosol. Optical Depth. Air pollution.

I. INTRODUCTION

The atmospheric pollution has become one of the most frequent topics on big discussions, showing that the resulting environmental impact is the biggest cause of the damages reflected on human health, ecosystems and materials.

The large urban centers progressively densify the population, supporting the intensification of anthropic activities. The emergence of industrial activities, the growth of vehicle fleets, fire, green areas devastation and deforestation, increasingly, damage the air quality.

The Federal District has, nowadays, a fleet of 1.321.552 motor vehicles [1]. The industries, represented by asphalt, cement, furniture, beverage, roasting and tire retread factories, also contribute with the emissions, but in smaller scale.

The Air Quality in the Federal District monitored by the Brasília Environmental Institute [3], which monitors 6 stations (Taguatinga, Rodoviária, Fercal I and II, CIPLAN and Queima Lençol), and the Interdisciplinary Center of Transport Studies (Centro Interdisciplinar de Estudos em Transportes - Ceftru/UnB) is in charge of 2 stations (714 south and HUB). The stations are fixed and only 3 parameters are quantified - out of 7 established by law [2]: Sulfur dioxide (SO_2), Total Particles in Suspension (PTS) and Smoke.

The monitored locations are considered critical sites, due to the intense motor vehicle traffic or to the existence of large cement factories. Both sources contribute to the high emission of elevated concentration of gaseous and particulate pollutants, which can be measured quantitatively at monitoring stations [3].

On previous studies about aerosol [4], it was shown that satellite data have been used to give information about air quality. The AOD can be analyzed by utilizing data from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor, which is an effective way to monitor and study the aerosol distribution and its effects along time, considering that the passing period is 1-2 days.

Therefore, this study’s purpose is to analyze the data to set up a multitemporal prospect of the aerosol behavior in the Federal District, from preexisting data of the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor name TERRA.

II. OBJECTIVES

The general objective of this study is to analyze the aerosol distribution in the Federal District between 2004 and 2013 using remote sensing techniques and geoprocessing. Therefore, the specific objectives are:

• To correlate the data obtained from the Air Monitoring Stations of the Federal District with the Aerosol product from Moderate Resolution Imaging Spectroradiometer MODIS sensor (TERRA) in the specific date collected from the stations;

• Multitemporal analysis (2004 to 2013) of the AE, which is the spectral dependency measure of the AOD.

ISBN: 978-1-61208-326-1

III. MATERIALS AND METHODS

The aerosol measurement in the Federal District does not exist (Figure 1). It is often limited to in situ analysis of Total Particles in Suspension (PTS) at the 8 air quality monitoring stations, and the information from the data are restricted to certain places.

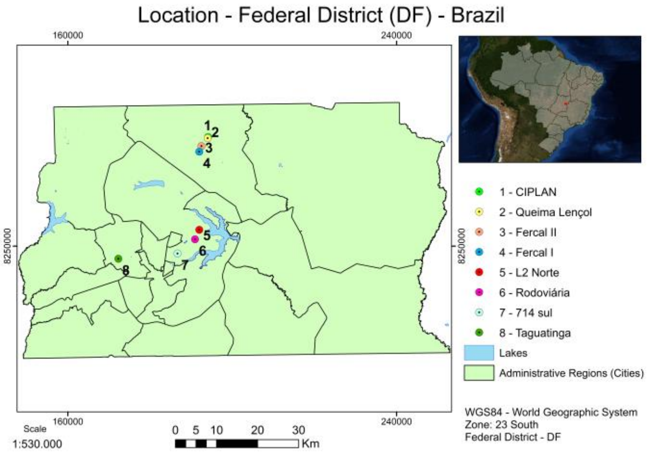


Figure 1. Location of the study area and the stations monitoring air quality.

The operating algorithm of the NASA (National Aeronautics and Space Administration) corresponds to the collection of aerosols group 5.1 (AQUA/TERRA Forward and Reprocessing) and provides for each pass daily files with values of the aerosol optical depth with spatial resolution of 10 km x 10 km over the ocean and the land [5].

Because of this limitation, to acquire the aerosol scenario in large scale encompassing the entire Federal District region, data from the MODIS's AERONET (Aerosol Robotic Network) [6] program is a federation of ground-based remote sensing aerosol networks established by NASA and PHOTONS (PHOtométrie pour le Traitement Opérationnel de Normalisation Satellitaire), ollaboration provides globally distributed observations of spectral aerosol optical depth (AOD), inversion products, and precipitable water in diverse aerosol regimes. Aerosol optical depth data are computed for three data quality levels: Level 1.0 (unscreened), Level 1.5 (cloud-screened), and Level 2.0 (cloud-screened and quality-assured). Inversions, precipitable water, and other AOD-dependent products are derived from these levels and may implement additional quality checks. Were selected (2003 to 2011), as it contains 36 bands, and the bands 1-2, 3-7 and 8-36 have special resolution of 250m, 500m and 1 km, respectively.

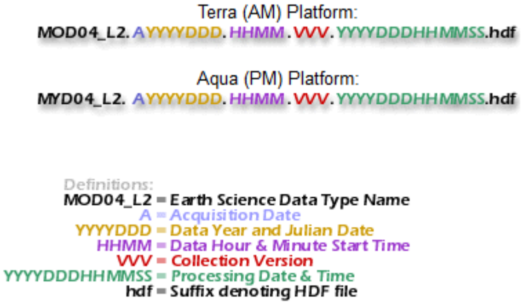
The data about aerosols were obtained from MODIS level 2, in type .hdf at NASA portal, at level 1 and atmospheric products data section, and are already processed and corrected. The date of the datum, the location, the period it was collected (day, night or both) and to which library it belongs (in this case, the choice was "Collection 5.1", which had all the data for the chosen time interval) are selected.

For this study, from 2010, daytime data was chosen, because the possible vectors with polluting potential have regular activity in this period. There were attempts to acquire products fortnightly, but it was concluded that there was not product available in the region in every pre-established date, so the acquisition was made in variable dates close to the interval pre-established.

In order to standardise the comparison of ground-based [7] observations with instant satellite-observed data at a given moment, the apparent reflectance and surface reflectance within an area of 10 km by 10 km centered at the Federal District observation station were averaged. Similarly, the ground-based data recorded within half an hour of the satellite overpass on cloud-free days were also averaged to derive AOD at the 550 nm wavelength using the Angström formula for validating satellite-based results [11].

AOD values from the high-resolution product land the standard Collection 5 AOD algorithm were compared to observed Aerosol Optical Thickness (AOT) values at three coastal Aerosol Robotic Network (AERONET).

These data has the nomenclature with the Julian data, specified as follow:



Terra (AM) Platform:
MOD04_L2.AYYYYDDD.HHMM.VVV.YYYYDDDHHMMSS.hdf

Aqua (PM) Platform:
MYD04_L2. AYYYYDDD.HHMM.VVV.YYYYDDDHHMMSS.hdf

Definitions:
MOD04_L2 = Earth Science Data Type Name
A = Acquisition Date
YYYYDDD = Data Year and Julian Date
HHMM = Data Hour & Minute Start Time
VVV = Collection Version
YYYYDDDHHMMSS = Processing Date & Time
hdf = Suffix denoting HDF file

Figure 2. Nomenclature model of the .hdf archives from MODIS.

At the processing of level 2 final data, the following softwares were used: ENVI (Plugin MODIS Conversion Toolkit (MCTK)), which is a module and does the conversion of sinusoidal projection to UTM and the ArcMap for the construction of the database for the maps of AOD, Corrected Optical Depth, Cloud Fraction, Mass Concentration and AE, which present the following parameters in the .hdf file:

- Corrected Optical Depth

(Corrected_Optical_Depth_Land)
Description: Optical Thickness corrected by 0.47, 0.55 and 0.66 μm
Valid interval: -0.05 to 5.0;

- AOD (Optical_Depth_Small_Land)

Description: Optical Thickness corrected for 0.47, 0.55, 0.66 and 2.13 μm
Valid interval: -0.05 to 5.0;

- Concentration Mass

(Mass_Concentration_Land)
Description: Concentration of mass on Earth.
Valid range: 0 to 1000 1.0e-6g/cm x ^ 2;

- AE

(Angström_Exponent_Land)

Description: AE from 0.47 to 0.67 µm
Valid interval: -1.0 to 5.0;

- Cloud Fraction

(Cloud_Fraction_Land)
Description: Cloud fraction from the aerosol cloud mask recovered and cloudy pixels, not including cirrus cloud mask.
Valid interval: 0.0 to 1.0.

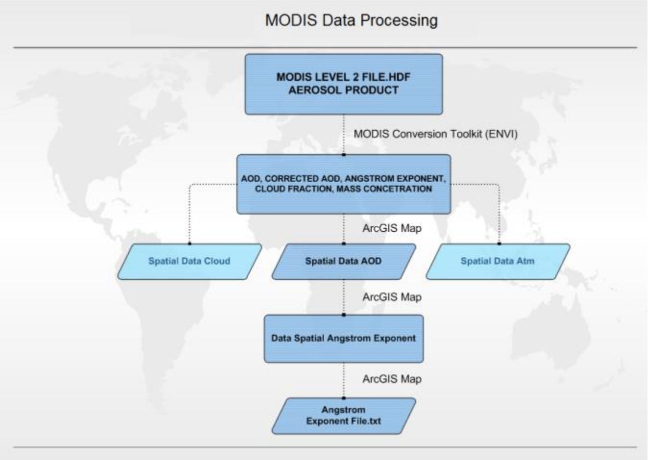

MODIS Data Processing
MODIS LEVEL 2 FILE.HDF AEROSOL PRODUCT
MODIS Conversion Toolkit (ENVI)
AOD, CORRECTED AOD, ANGSTROM EXPONENT, CLOUD FRACTION, MASS CONCETRATION
ArcGIS Map
Spatial Data Cloud
Spatial Data AOD
Spatial Data Atm
ArcGIS Map
Data Spatial Angstrom Exponent
ArcGIS Map
Angstrom Exponent File.txt

Figure 3. Stage of the data processing from Level 2 MODIS extraction to the final analysis phase using environment ArcGIS.

The first 19 bands are positioned in the region of the spectrum eletroradiometric situated between 405 nm and 2155 nm, so that the bands are 1-7 targeted for terrestrial applications; bands 8-16 for ocean observations and bands 17-19 for atmospheric measurements [12].

The sampled pixel has a 10 km spatial resolution, because data with smaller spatial resolution (ex: 250 m) were not available for the region. The histogram was assembled on ArcMap and the graphic maps of the above described items were made available.

The Interdisciplinary Center of Transport Studies (Centro Interdisciplinar de Estudos em Transportes - Ceftru/UnB) provided the spread sheets with the data of the Air Quality Monitoring Stations from the 8 stations located in the Federal District.

The *shapefiles* of the Census Sector 2010 [9] were used to verify the influence of the urban areas on the distribution of aerosol at the regions of the Federal District, Hydrography and Road System, from Real Estate Company of Brasília (Terracap) origin (2010) for region localization reference.

All provided and generated data can be found in UTM SIRGAS 2000 (Geocentric Reference System for the Americas), zone 23 south.

IV. Results and Discussion

From 2003 to 2009, there was not daily accurate information that could be correlated with the data from the MODIS aerosol product. However, the Ceftru collected significative data for the year 2010, from the 8 stations, from which two close dates were selected, when there were data about aerosol in the Federal District area (10/04/2010 and 10/14/2010), during rainy season; there were not data production during the dry season (Tables I and II).

TABLE I – TOTAL PARTICLES IN SUSPENSION ON 04/10/2010.

04/10/2010		
ID	Station	PTS (µg/m³)
1	CIPLAN	1.250,06
2	Queima Lençol	-
3	Fercal II	-
4	Fecal I	218,05
5	L2 Norte (HUB)	96,92
6	Rodoviária (central bus station)	148,01
7	714 Sul	66,45
8	Taguatinga	240,98

On the data shown for these days, the average standard of AOD is below 0,5 µm. Figures 11 and 16 show the level of particle size variability that is affected by the sum of the types of activities in the Federal District. There are many activities that can influence a region, where a pollution with different particle sizes will arise. The result of the AE points the difference between the values of AE for the dry and rainy seasons. Previous studies [6] mention that the tendency is that during the rainy season, the value can be up to 1 µm, in the case of the analyzed dates, varying from 1,35 µm to 1,92 µm, which means that the type of aerosol is compound of thin particle. Usually, on a rainy day, the aerosol goes down to the ground surface, due to the entrainment by the water drop.

On 10/04/2010, the data from the air quality monitoring stations show a substantial particle concentration at the CIPLAN station (1.250,06 µg/m³), where the AE is 1,92 µm (Table I and Figure 6). It is known that the higher its value, the bigger the spectral dependency, which means bigger optical depth variation with the wavelength; in other words, the AE is related to the medium size of the aerosol particles on the atmospheric column, since smaller particles has bigger spectral dependency that larger ones.

Some studies [4] in the dry season also showed that soil dust from agricultural land un-vegetated being great source of aerosols in regions with arid characteristics. AOD is closely related to the topography [10], suggesting that a significant inverse correlation exists between them. AOD is significantly lower in high-altitude areas. Conversely. anthropogenic activities produce large amounts of fine aerosol particles in densely populated urban areas. However, airborne dust originating from traffic and construction activities in urban areas, in addition to larger particle sizes of soot aerosols produced by industrial and civil coal fuel combustion cause coarse-mode aerosols.

The particles originated in this region activities are usually from mechanical and chemical processes. The latter are, generally, smaller, unlike the ones made during the production of dust by the cement factories. So, the AE is also often used as an indicative of regions where aerosols of different types are predominant, originated from several processes.

The same situation happens at the Taguatinga station, where there is a large urban center that produces aerosols resulting from chemical processes like engine combustion. The AE value is 1,35 μm (Figures 6 and 10), and reduces to 0,52 μm on 10/14/2010, which presents in the zone of aerosol made by thin particles, with less spectral dependency.

This tendency to smaller values can be linked to a station where there is a “renovation” of the air column because of rain and wind, decreasing the values for regions with an expected polluting potential. Both dates show similar scenarios, without a very significative variation of the data of optical depth and AE. Only the spatial distribution of the datum varied.

Similar fact happened to the mass concentration parameter, which is the aerosol column on the atmosphere. It varied up to 1,35 μm on 10/14/2010, and also the spatial distribution in the region varied.

The figures (Figures 4, 6, 8 and 10) below show the AOD and AE for the dates of 04/10/2010 and 14/10/2010, which could be directly correlated with the existing data in fixed stations stations.

TABLE II – TOTAL PARTICLES IN SUSPENSION ON 14/10/2010.

14/10/2010		
ID	Station	PTS (μg/m³)
1	CIPLAN	971,71
2	Queima Lençol	-
3	Fercal II	-
4	Fercal I	246,07
5	L2 North (HUB)	72,73
6	Rodoviária (central bus station)	114,76
7	714 South (714 sul)	66,45
8	Taguatinga	208,64

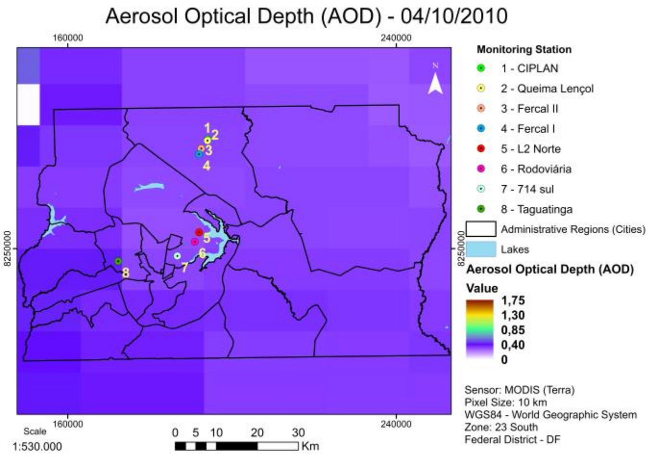


Figure 4. AOD for the day 04/10/2010, indicating the location of the monitoring stations and the population distribution region.

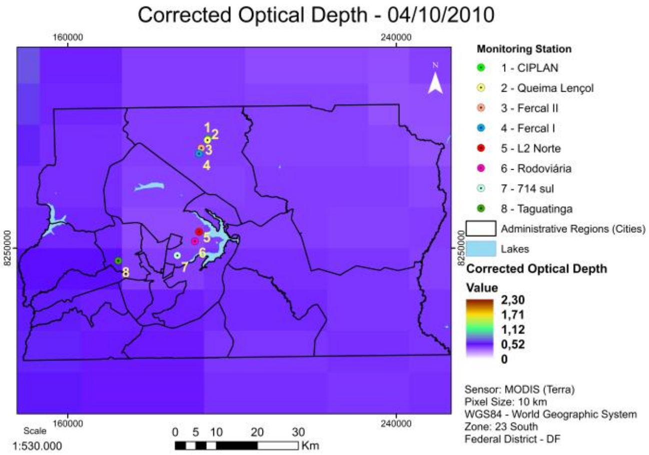


Figure 5. Corrected Optical Depth for the day 04/10/2010, indicating the location of the monitoring stations and the population distribution region.

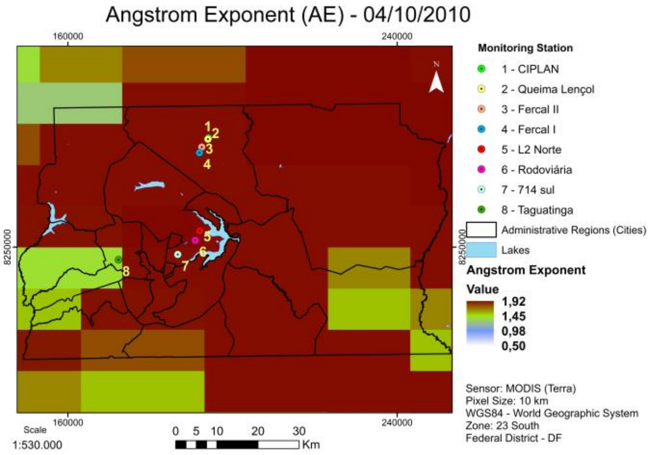


Figure 6. AE for the day 04/10/2010, indicating the location of the monitoring stations and the population distribution region

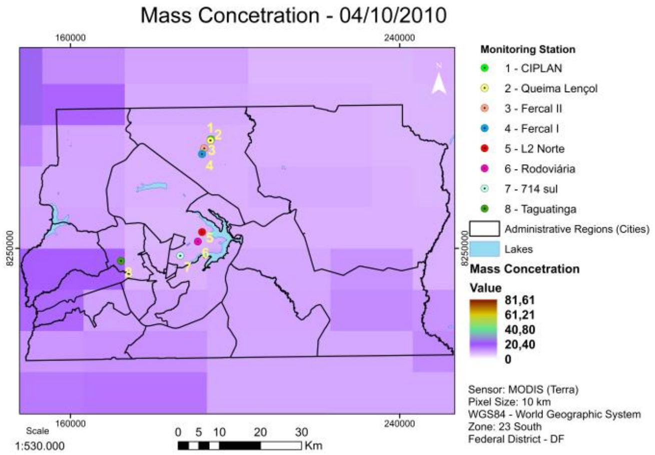


Figure 7. Mass concentration for the day 04/10/2010, indicating the location of the monitoring stations and the population distribution region

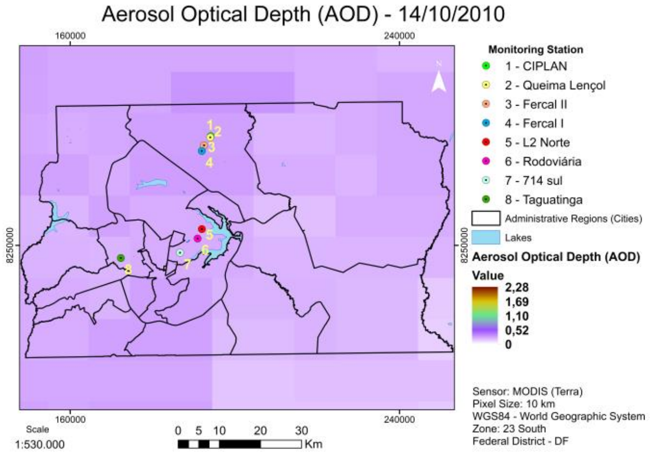


Figure 8. AOD for the day 14/10/2010, indicating the location of the monitoring stations and the population distribution region.

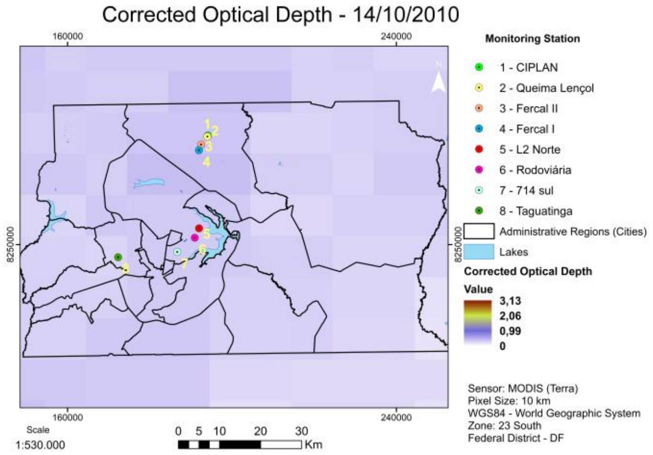


Figure 9. Corrected Optical Depth for the day 14/10/2010, indicating the location of the monitoring stations and the population distribution region.

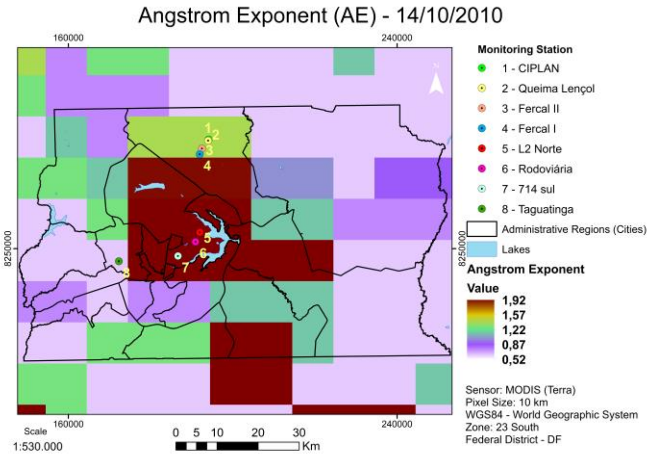


Figure 10. AE for the day 14/10/2010, indicating the location of the monitoring stations and the population distribution region.

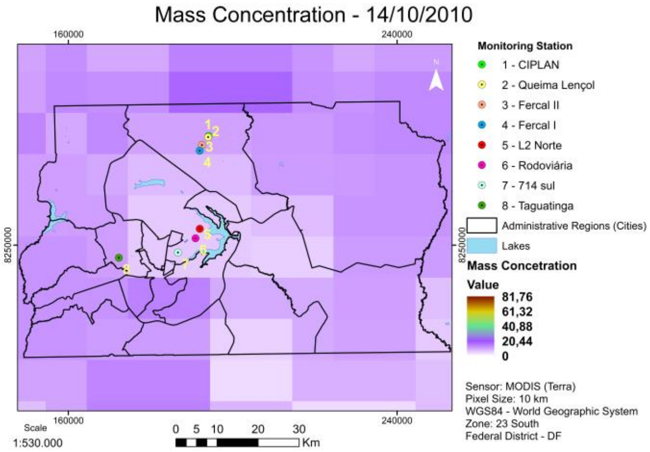


Figure 11. Mass concentration for the day 14/10/2010, indicating the location of the monitoring stations and the population distribution region.

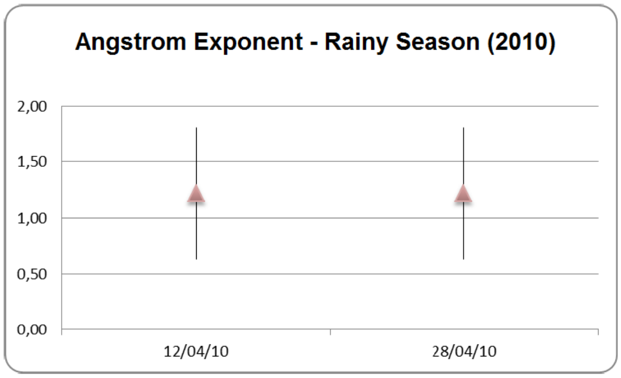


Figure 12. AE - Rainy Season – (only data MODIS) (2010).

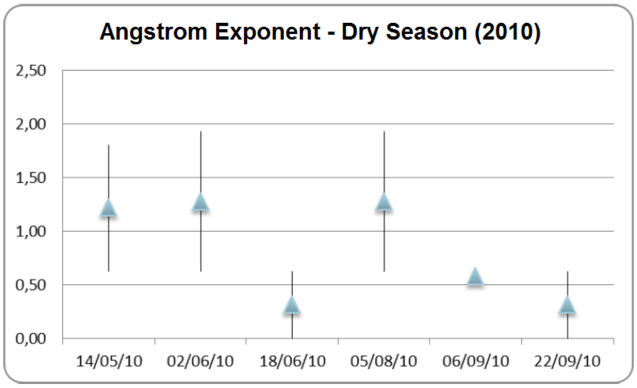


Figure 13. AE – Dry Season (only data MODIS) (2010).

Overall, during the dry season, the large particles stayed on the atmosphere longer than during the rainy season, because during the drought there is no process of "rinsing" the air, proving in most parts of the analysis that, by the distribution of the value of the AE in the region, the pixel average value was ≤ 0,5, which is equivalent to larger particle sizes.

The downtown area of Brasília and the surrounding of the Fercal area, in general, were the ones in the scenario that maintained most part of the concentration of AOD and the highest values of the AE, along the years. It can be attributed to the types of activities existing in these regions. Taguatinga showed a smaller variation of the AE and bigger AOD variations in some dry season situations.

The lack of more recent data (from year 2011) prevented subsequent correlation to aerosol data from MODIS with ground stations (the stations are government and have no more maintenance).

V. Conclusion And future work

The range of the AE analyzed indicates the level of the particle variability that can be correlated with the types of activities of a specific region.

The particle sizes in urban areas tend to be larger compared to the ones found in natural environments, also during drought, when the distribution of the particle sizes tends to be more varied. Sand and dust were the main sources of AOD in Federal District.

AOD was affected mainly by human activities and showed a slight increasing trend. Anthropogenic activities caused the increase of fine mode AOD in most areas. In addition, variation in fine mode aerosols dominated the yearly fluctuation of AOD, and the main aerosol type shifted gradually to the urban industrial type.

During the rainy season (Figure 12), the particles tend to be smaller, with the average AE between 0,5 to 2, while the particle size during the dry season (Figure 13) tend to be more varied, as mentioned in previous studies [4][5].

Overall, the areas with industrial activities are the main source of pollution, as well as big urban centers.

It is noteworthy that more accurate field validations, with a larger number of parameters at the existing air monitoring stations would make it closer to the reality of the information monitored by the MODIS satellite, leading to a safer correlation with the final result of an analysis of this kind.

Lesson learned and indication to future studies with radiometers and spectrometers will help understand the particles, specifying the pollution type existing in each region, thus, enabling to compare which types of sources cooperate with the dispersion of aerosol in the Federal District.

In future work we intend to use low-cost sensors in mobile devices for increased accuracy in generating data on ground level for correlation with data from MODIS.

References

[1] Denatram. Car fleet – 2011 [Online]. Available online: http://www.denatran.gov.br/frota.htm (acessed on 2014.02.15).

[2] Ibram. Monitoring Report of Environmental Quality Air Quality in Federal District - Brasilia Environmental Institute, Brasilia, 2008.

[3] Conama - National Council for the Environment, Conama Resolution 03. Establishes standards for air quality. Brasilia, 1990.

[4] Rosida. Variability of AOD Over Java Continent by Using MODIS Data. Applications of Climatology And Environment Division, National Institute Of Aeronautics And Space, 2007.

[5] Lorraine Remer, Yoram Kaufman, and Didier Tanré,. Algorithm for remote sensing of tropospheric aerosol from MODIS: Collection 005 (Product ID: MOD04/MYD04). NASA, 2006.

[6] ZiPeng Dong, Xing Yu, XingMin Li, and Jin Dai. Analysis of variation trends and causes of aerosol optical depth in Shaanxi Province using MODIS data. Meteorological Institute of Shaanxi Province - China, 2013.

[7] Aerosol Robotic Network (AERONET). Available online: http://aeronet.gsfc.nasa.gov/ (accessed on 2014.02.10).

[8] Junliang He, Yong Zha, Jiahua Zhang , and Jay Gao. Aerosol Indices Derived from MODIS Data for Indicating Aerosol-Induced Air Pollution. Laboratory of Virtual Geographic Environment, Ministry of Education, College of Geographic Science, Nanjing Normal University, Nanjing, China, 2014.

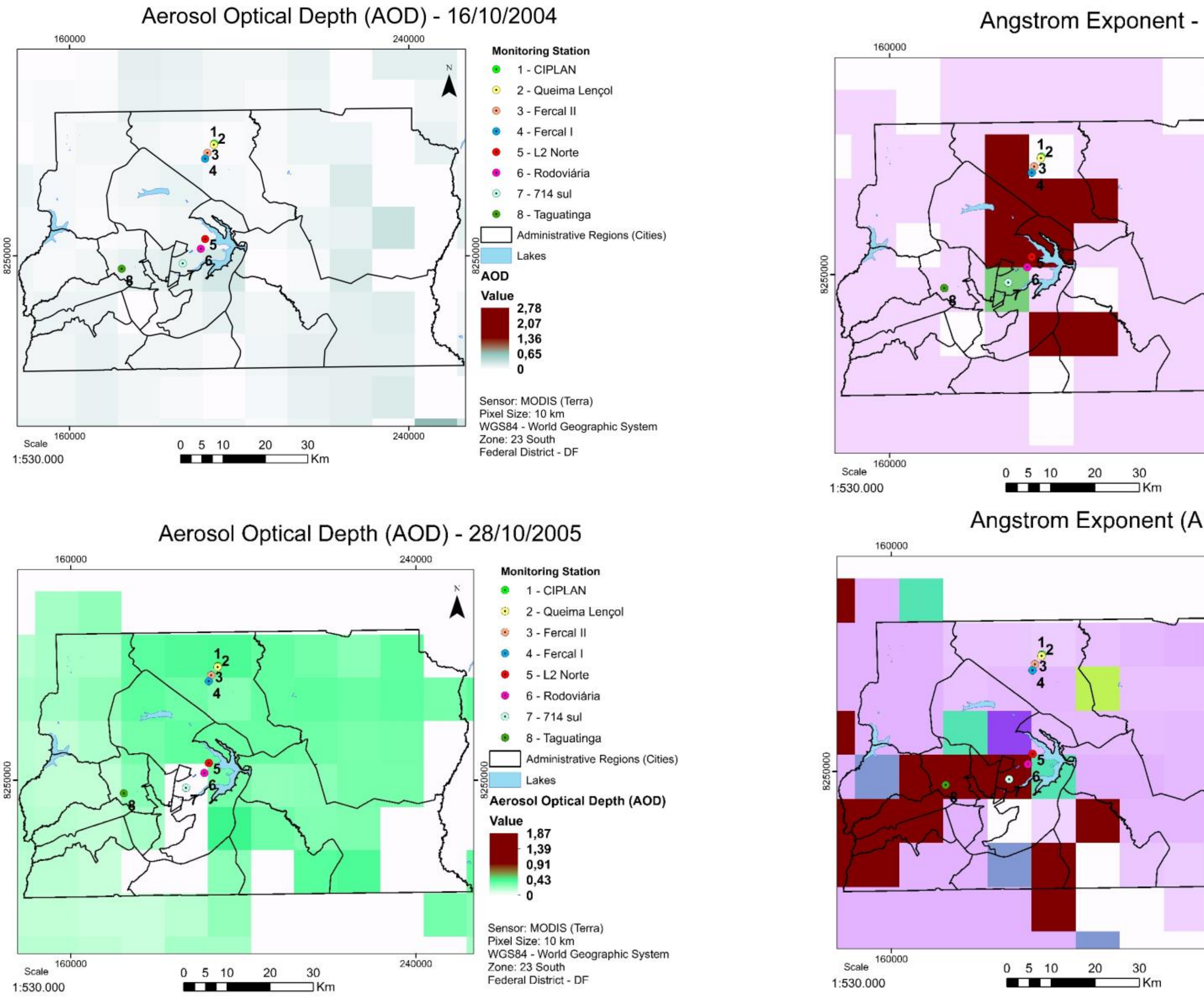
[9] Ibge - Brazilian Institute of Geography and Statistics. 2010 Census. Brazil, 2010.

[10] Anup Prasad, Ramesh Singh, and Ashbindu Singh. Variability of Aerosol Optical Depth Over Indian Subcontinent Using MODIS Data. Journal of Indian Society of Remote Sensing vol. 32, no. 4. December, 2004.

[11] Ångström, A. The parameters of atmospheric turbidity. Tellus 1964, 16, 64–75.

[12] John Barker, Paul Anuta, and Joann Harnden. MODIS spectral sensivity study: requirements and characterization. Washington: Nasa, Oct, 1992, 84p.

ATTACHMENTS



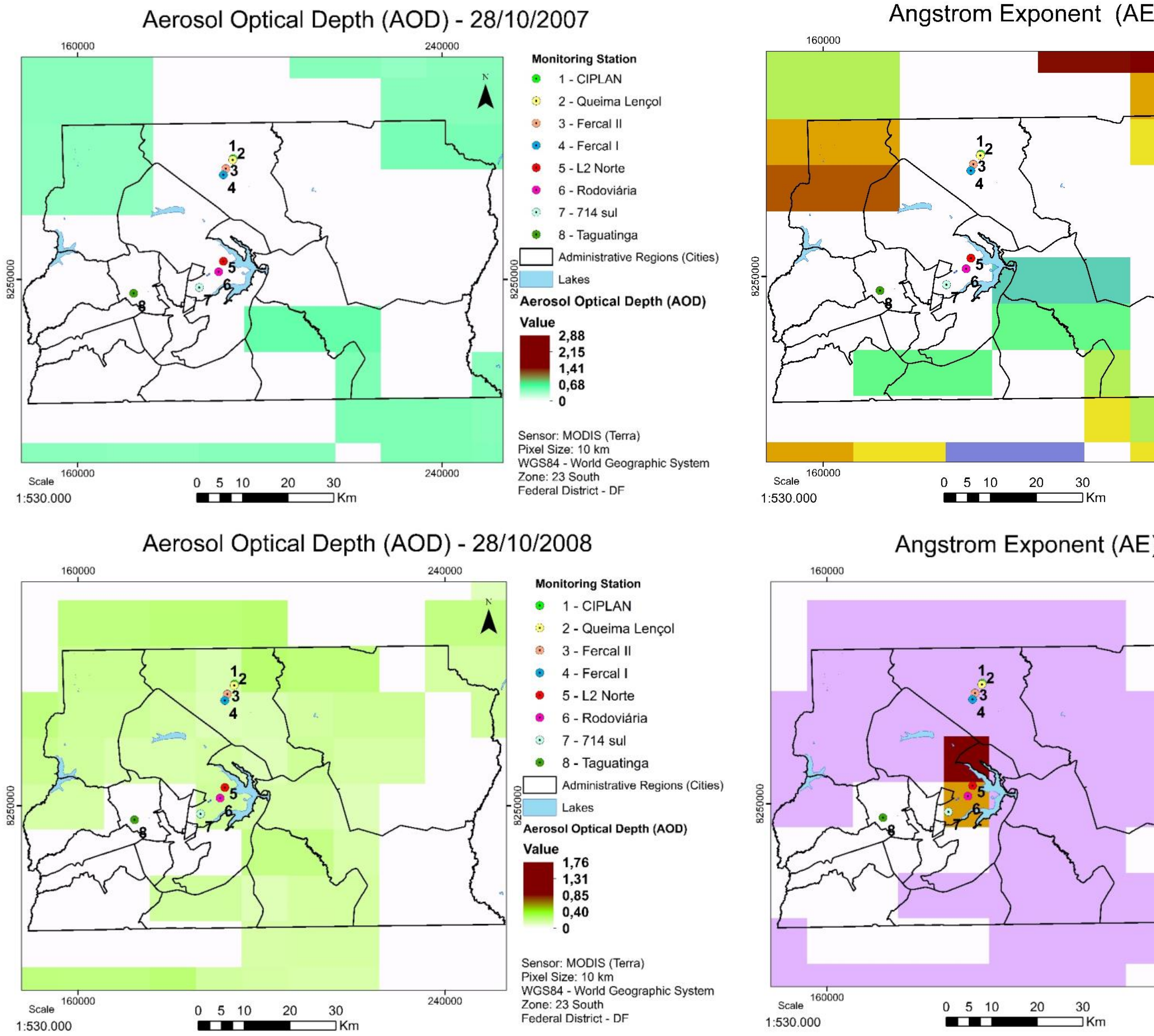
Aerosol Optical Depth (AOD) - 16/10/2004

Angstrom Exponent -

Aerosol Optical Depth (AOD) - 28/10/2005

Angstrom Exponent (A

Figure 14. AOD and AE for the days 16/10/2004 and 28/10/2005.



Aerosol Optical Depth (AOD) - 28/10/2007

Angstrom Exponent (AE

Aerosol Optical Depth (AOD) - 28/10/2008

Angstrom Exponent (AE)

Figure 15. AOD and AE for the days 28/10/2007 and 28/10/2008.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

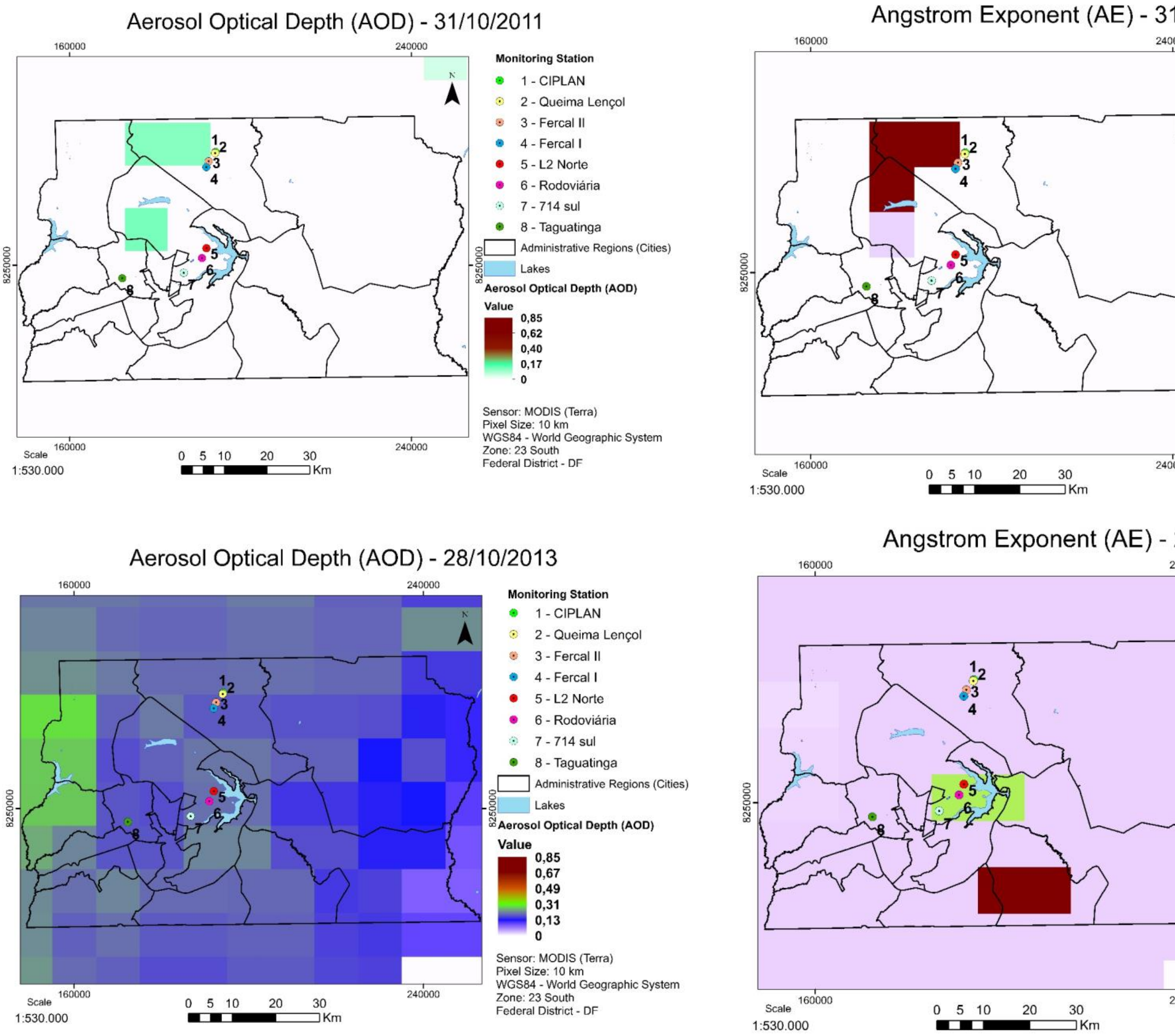


Figure 16. AOD and AE for the days 31/10/2011 and 28/10/2013.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

Evaluating Principal Components Analysis of Particular Spatial Statistical Models

Mauro Mazzei, Armando Luigi Palma
National Research Council
Institute of Systems Analysis and Computer Science
Rome, Italy
e-mail: mauro.mazzei@iasi.cnr.it, palma@arpal.it

***Abstract* — This work is based on an analysis of the main components derived from particular patterns of spatial statistical data. The reference models of spatial statistical analysis are extracted only from the data of bi-temporal aerial photographs. This methodological approach introduces a significant improvement in the evaluation of changes in the territorial scenery, providing a wider interpretation of the problems of the area studied and encouraging a more analytical reading of complex environmental phenomena. In order to improve reading and analysis of the territorial changes it is necessary to compare the same geographical space in two different moments that enclose a well-defined period of time.**

Keywords – GIS; cartography; spatial data analysis; spatial statistical model.

I. INTRODUCTION

The techniques of analysis of multi-temporal remote sensing images are currently based on the recognition of the diversity of spectral indices of the two images observed. This methodology is applied by using software products that classify the content of the images. This classification is based on the similarity of the local spectral radiance. Classes of pixels are classified by evaluating the evolution of the state, from beginning to end [1]. The classification is made on the basis of spectral responses of the surfaces, based on the concept of similarity between pixels. This clustering is applied to the pixels of each image. This methodology requires the definition of prototypes based on a comparison of a minimum set of pixels representative of the class. There are basically two different methods for automatic classification of digital images [2]:

- Supervised Classification: this is used for a quantitative analysis of remote sensing images. This methodology can be summarized in two phases. The first phase provides a definition of the legend specifying the set of cover. The second phase provides an identification of the spectral signature on the ground. These two phases are useful for providing information on the software used to carry out the classification of the area [3][4].
- Unsupervised Classification: this uses the concept that produces similar spectral responses. A specific knowledge about the extent, type, and class descriptions is not required with these systems. These properties are based on observations of clusters formed in the spatial features. The main peculiarity is that the classes are identified by cluster compact and it is easy to distinguish between them. These clusters do not require knowledge of the features or of their nature. The entire space is split into spectral classes according to criteria of proximity or similarity. Only after the process of identification of spectral classes is finished, will the analyst associate these properties in relation to the knowledge of the territory [5].

In this paper, we propose a method that belongs to the techniques of unsupervised classification. With each cluster, in addition to the spectral properties, proximity and similarity, there are also recognized: perimeter, area, their relationship, their moments of inertia Jx and Jy, etc. Each cluster extracted from the digital image is described by its radiometric, geometric, and inertial properties, which are stored in a database. Applying the principal component method [9] to clusters extracted from digital images and described by their properties, as mentioned above allows us to identify, for each of the new factorial axes, similar classes of cluster characterized predominantly by only variables that have a high correlation value (factor loadings) between variable and factor.

The clusters, in the new reference by each main component, retain the property of having an average equal to 0 and variance equal to 1. The graphic rendering, for each main component of the clusters of positive coordinates, allows the segmentation of the original image. This shows classes of clusters that often are not detectable in the original image; therefore, the comparison between the segments of multi-temporal images obtained in this manner is facilitated and more readily available.

The paper is structured as follows. In Section 2, we illustrate the location of the study area, the material used, the organization of the digital data collection and preparation of digital data. In Section 3, we describe the method for the analysis of the data. In Section 4, we describe the type of statistical analysis used. In Section 5, we propose the data model used for the analysis, we provide examples in order to show our proposal. Finally, in Section 6, the discussion and conclusion are given.

II. DATA ORGANIZATION

The area of study is located in the south of Italy, in the province of Taranto, and is located in the northern Gulf of Taranto within the first basin of the Little Sea within which flows the river Galeso.

The material used for the analysis of this data is in raster digital format, with reference to an aerial photo of the Military Geographical Institute of 1940. The analyses were made available in digital format thanks to the collaboration of the Laboratory of ancient topography and photogrammetry of the University of Salento. The Geoportal's Web Map Service (WMS) of the Campania Region, which has kindly allowed us to use the orthophotos of the same site obnserved in 2010, as well as a portion of a topographic map in scale 1:10.000, used for georeferencing of both aerial photos.

The raster digital format was acquired through a photogrammetric scanner; georeferencing [6][7], was then used. The system used is the Universal Transverse Mercator (UTM) - European Datum 1950 (ED50).

The georeferencing of the Regional Technical Map requires a value of Root Mean Square (RMS); the graphics below show the graphical representation error with respect to the scale representation (1:10000 scale is 0.2 mm, 2 meters in real scale) [6][7].

This technique allows us to identify homologous points in the aerial photo for the georeferencing of the same space, without resorting to tools of GPS positioning, in order to detect in situ. In this way we obtain a minimum margin of error for the measurements of the spatial analysis for the case study.

The recognition of homologous points between technical mapping and aerial photo requires an accuracy for the identification of the exact location and a good distribution of support points; in this scenario, five control points are identified, which gave good results for a correct georeferencing of the aerial photo.

The transform algorithm applied to the matrix is of the type polynomial affine [6].

The recognition of homologous points in the aerial photo of 1940 is more difficult, since the change in territory does not permit an easy identification of common elements with aerial photos of the Military Geographical Institute.

In this case, we first identify large objects such as the geomorphological aspects, then small elements, for example, buildings and any other element which allows a good distribution of the points of support for the application of the georeferencing method.

The picture of 1940 is not a calibrated aerial photo and therefore the picture is distorted and requires an appropriate transformation. A transformation involves a mapping of locations of points in one image to new locations in another. Image Transformations used to align two images may be global or local. A global transformation is given by a single equation which maps the entire image. Examples are the affine, projective, perspective, and polynomial transformations. Local transformations map the image differently depending on the spatial location and are thus much more difficult to express succinctly. In the image of 1940 a local transformation was used [6].

The Pixels of the aerial photos positioned in their correct geographic space have been a simply of the nearest/neighbor interpolation.

III. DATA ANALYSIS

The two orthophotos from Figure 1 and 2 show the Northern arc of the first Sinus of the Mar Piccolo, into which flows the river Galeso, dating back to 1940 and 2010, respectively.

An initial examination of comparative urbanization can be seen in this frame of 2010, which mainly affected the western arc of the first Sinus of the Mar Piccolo. Variations seem to be able to detect an industrial plant related to “Cantieri Navali”, located north of the Sinus, and then, along the course of the river that flows near Galeso, the same shipyards that are no longer used.

Our curiosity extended far beyond a first visual comparison of the two images from which you can detect, even with the naked eye, the profound changes that occurred over a period of seventy years, especially in the first western Sinus of the Mar Piccolo.

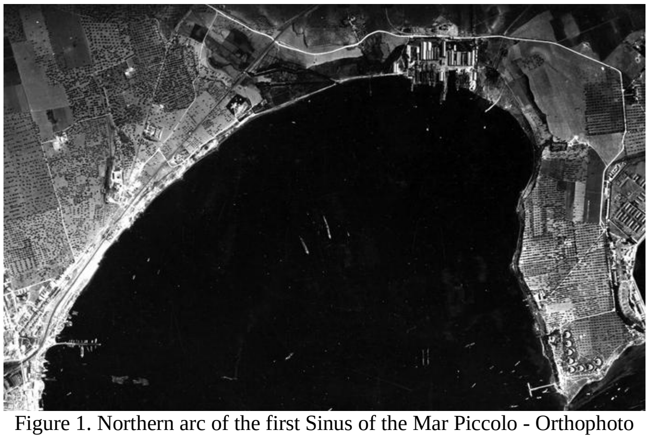


Figure 1. Northern arc of the first Sinus of the Mar Piccolo - Orthophoto Institute Geographical Military (IGM) 1940

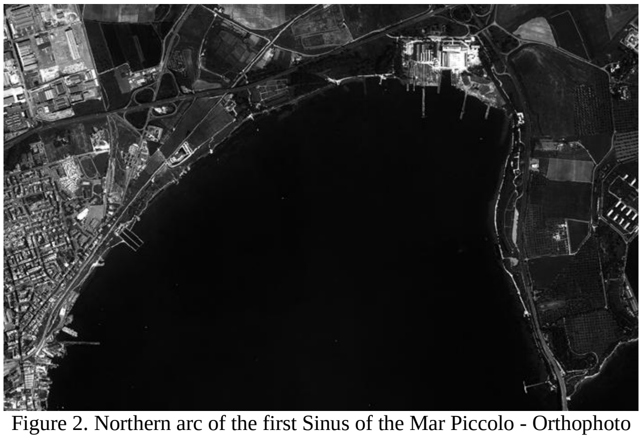


Figure 2. Northern arc of the first Sinus of the Mar Piccolo - Orthophoto Web Map Service (WMS) 2010

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

We, therefore, subjected the two images, in view of their automated processing, with a scanning step of 20 microns, obtaining their digitization by a mosaic in which each tile was represented by its coordinates x, y and by the value of 8 bits of its gray level.

For each image, the mosaic thus obtained, we proceeded with an automatic extraction of “patch” with the criterion of “similarity” between tiles assigned to satisfying the criteria of closeness and proximity of the relative levels of gray.

“Patch” in ecology means the structural unit of an environmental system heterogeneous, identified on the basis on the differences that appear within the area itself.

We have thus converted each image into a set of “patches” each of which were calculated fifteen numeric attributes related to the geometric properties of the patch (perimeter, delta-X and delta-Y, area, etc.) and others to its inertial properties (Jx, Jy, Rx, Ry, etc.).

From the image of 1940 were extracted 341 patches, whereas 438 will be extracted from the image of 2010, about a third more of the patches contained in the image of 1940. This first result shows how the environment has increased the fragmentation of the examined area.

The two analog images, when converted into two distinct sets of patches, each characterized by fifteen attributes, were subjected to a factor analysis procedure with the method of principal components (Hotelling) [8][9][11].

The Principal Components Analysis (PCA) is a linear transformation that transforms the data into a new coordinate system; the new set of variables, the principal components, are linear functions of the original variables and are uncorrelated. The greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. In practice, this is achieved by computing the covariance matrix for the full data set. Next, the eigenvectors and eigenvalues of the covariance matrix are computed, and sorted according to decreasing eigenvalue. One can see that the PCA's bias is not always appropriate; features with low variance might actually have high predictive relevance; this depends on the application [10].

Given a set of *p* observations for each variable of a complex of *m* variables, the principal component analysis is proposed to determine new variables linearly related with the given variables, but in a lesser number of these latter, so that we can represent the variability expressed by the original variables. If it is not possible to meet these conditions, it is not possible, to represent the variability of the original variables with less than *m*, t and he principal component analysis is limited to an acceptable extent represent the majority of this variability with less than *m* variables. The problem of the analysis of the components is therefore related to the reduction of the number of descriptive variables *m* of *p* objects, regardless of the ability to identify new variables; Such identification must be decided in each particular case, generally, without any reference to the statistics, and in the field of the phenomena involved in the study [12][13].

IV. STATISTICAL ANALYSIS

The evaluation of the variance between the two aerial photos examined is based on the classification of patches extracted and subjected to statistical methodology of the analysis of the main components (Hotelling). The principal component analysis is a multivariate statistical technique that explains the variability of a statistical variable in k dimensions $Z=(Z_1,Z_2, \ldots, Z_k)$ in terms of k variables $Y_1,Y_2,\ldots Y_k$, linear combinations of the Z_j. It has:

$$Y_i = \sum_j b_{ij} Z_j \ (i=1,2,\ldots,k) \qquad (1)$$

where b_{ij} are constants to be determined. Y_i are called the main components of the variable Z and assuming they are not related to each other ordered by importance, in the explanation of the variability of Z we have:

$$\mathrm{cov}(Y_i,Y_j)=0 \ \ (i \neq j) \qquad (2)$$

$$V(Y_1) \geq V(Y_2) \geq \ldots \geq V(Y_k) \qquad (3)$$

where *cov* is covariance and *V* is variance. Without loss of generality we can assume that the variables Z_i are standardized, with mean equal to 0 and variance equal to 1, so as to eliminate the influence of the origin and the unit of measurement data, so that it results the following expression:

$$Z_j = (X_j - \mu_j)/\sigma_j \qquad (4)$$

Also, impose the condition that the overall variance of Z_j is equal to that of Y_i, i.e.:

$$\Sigma_i V(Y_i) = \Sigma_i V(Z_i) = k \qquad (5)$$

At last, suppose that the vectors

$$b_i = (b_{i,1}, b_{i,2},\ldots, b_{i,k}) \qquad (6)$$

have unit length, i.e., they fulfill the condition:

$$\Sigma_j b^2_{ij} = 1 \ (i=1,2,\ldots, k) \qquad (7)$$

On account of this, the vectors b_i that maximize the variance of Y_1, of Y_2, …, to Y_k with the constraints (3) and (4), are the eigenvectors of the matrix C of the coefficients of correlation between the variables Z_j, which correspond to

ISBN: 978-1-61208-326-1

the eigenvalues λ_1, λ_2,…, λ_k of **C**, sorted by non-increasing value. We then have:

$$|C - \lambda I| = 0 \qquad (8)$$

$$b_i (C - \lambda_i I) = 0 \qquad (9)$$

where I is the unit matrix. The matrix C is symmetric and positive definite for which the solutions λ_i of the (8) are non-negative and such that their sum (trace of the matrix C) is equal to k. We then have:

$$\Sigma_i \lambda_i = k \quad (i=1,2,\ldots, k) \qquad (10)$$

The variance of the i-th component is:

$$V(Y_i) = \lambda_i \qquad (11)$$

And the contribution of Y_i to the overall variance is:

$$P_i = V(Y_i) / k = \lambda_i / k \qquad (12)$$

V. Spatial Statistical Models

This procedure allowed us to calculate the correlation coefficients between the 15 variables adopted to describe each patch, with the aim of drastically reducing the number of variables, thereby explaining the overall variability of the system with a smaller number of attributes, each of which appears to be a linear combination of the attributes of departure.

TABLE I. VARIABLES OF THE MODEL USED

Variable	**Description**
Nz	Z coordinate - Average depth
Nt	N.ro of pixels of the object
Area	Attributes of the object:-Area
Perimeter	Perimeter of the object (Edge detection – Sobel)
DeltaX	Xmx-Xmn
DeltaY	Ymx-Ymn
IdealArea	Ideal Area = DeltaX * DeltaY
Gx	- barycentre X
Gy	- barycentre Y
Jx	- moment of inertia with respect to the X axis
Jy	- moment of inertia with respect to the Y axis
Rx	- radius of inertia X
Ry	- radius of inertia Y
AreaRect	- area of the circumscribed rectangle.
RapportAAR	- relationship between area and area of the circumscribed rectangle.

By using the factor analysis of the first array of data - about the image of 1940 – there emerged five factors that explained 96% of the total variance of the system. Of these five factors, the first alone explained 49% of the variance while the other four factors explained, neatly, 22%, 13%, 7% and 5%.

Table II reports the values of composition of each factor in function of the original variables. The weights of the variables were calculated on each factor (factor loadings).

These weights may also be interpreted as the correlation coefficients between variable and factor.

TABLE II. ROTATED FACTOR MATRIX (FACTOR LOADINGS) - IGM 1940 - NUMBER FACTOR 5

Weight of the variables	**F1**	**F2**	**F3**	**F4**	**F5**
Nzm is correlated with factor 3	0.3294	-0.5032	0.7269	-0.1063	-0.1220
Nt is correlated with factor 2	0.0225	0.9623	0.0025	-0.0441	0.2119
Area is correlated with factor 2	0.0225	0.9623	0.0025	-0.0441	0.2119
Perimetro is correlated with factor 1	-0.8277	0.1478	-0.4097	0.1351	-0.0239
DeltaX is correlated with factor 1	-0.9129	0.2140	-0.0695	-0.2047	0.2345
DeltaY is correlated with factor 1	-0.7447	0.1449	0.2987	0.1453	0.4479
AreaIdeale is correlated with factor 1	-0.9084	0.2367	-0.0920	-0.2057	0.2198
Gx is correlated with factor 4	0.0999	0.0407	0.1019	0.9694	0.1508
Gy is correlated with factor 3	0.2159	0.0369	0.9192	0.2047	0.1815
Jx is correlated with factor 2	-0.3072	0.9288	-0.1034	0.0636	-0.0637
Jy is correlated with factor 2	-0.3076	0.9286	-0.1037	0.0636	-0.0635
Rx is correlated with factor 1	-0.9788	0.0787	-0.1247	-0.0472	0.0154
Ry is correlated with factor 1	-0.9795	0.0756	-0.1216	-0.0396	0.0155
AreaRett is correlated with factor 1	-0.9475	0.0306	-0.1927	0.0140	-0.0814
RapportoAAR is correlated with factor 5	-0.1775	0.1713	0.0751	0.1507	0.9249

We reduced the size of the area of the patch definition, the image of 1940, from 15 to 5, after which we reconstructed the images with the new standardized patch values greater than average, that is greater than zero - based on each major component. These are illustrated in Figure 3, Figure 4, Figure 5, Figure 6 and Figure 7.

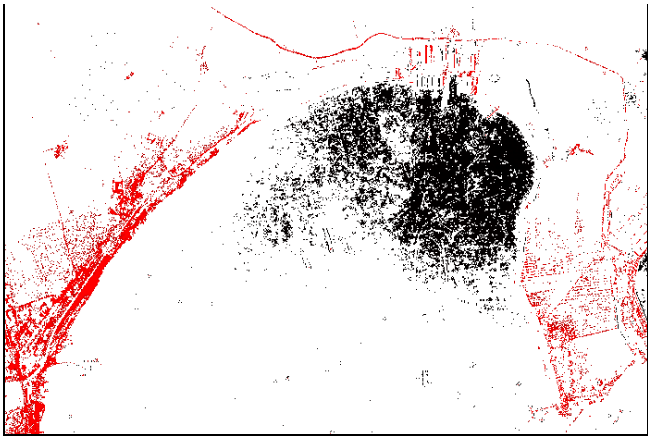


Figure 3. The reconstructed image with the I main component – 1940

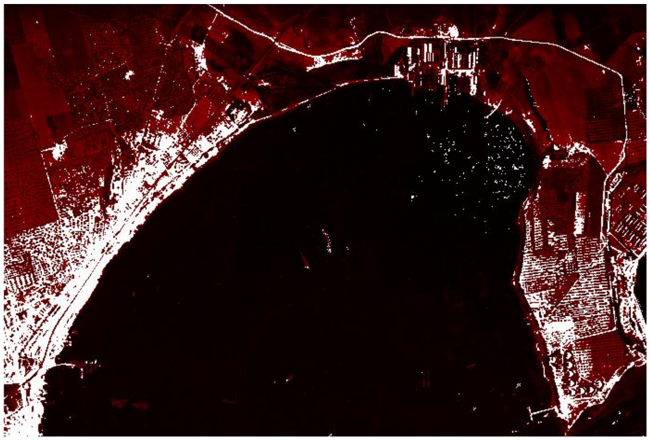


Figure 4. The reconstructed image with the II main component – 1940

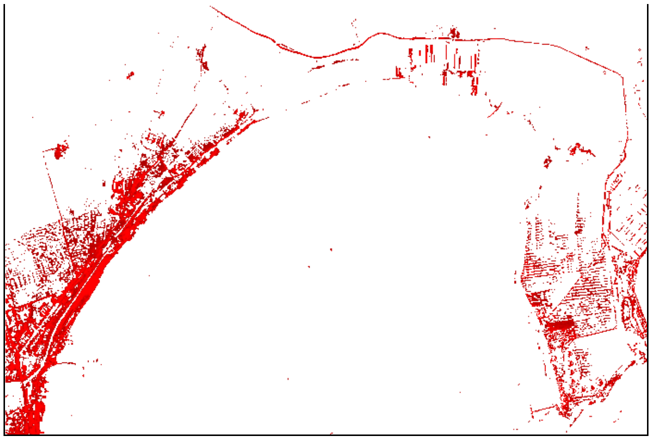


Figure 5. The reconstructed image with the III main component – 1940

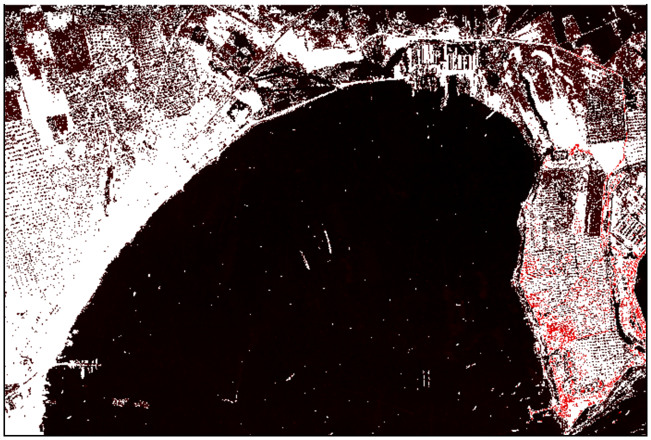


Figure 6. The reconstructed image with the IV main component – 1940

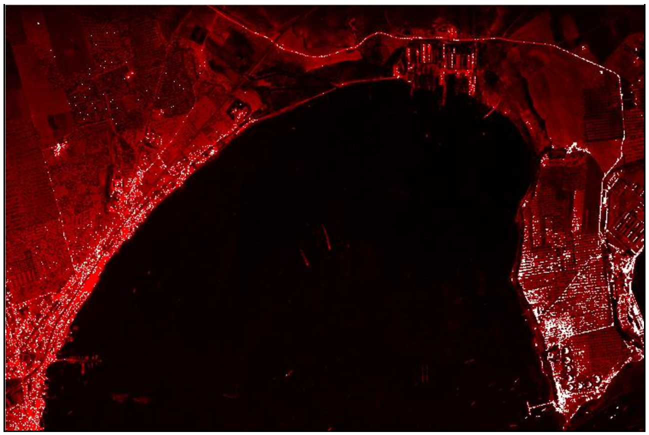


Figure 7. The reconstructed image with the V main component – 1940

Similarly, we proceeded to the WMS image taken in 2010. From this processing four main components (or factors) emerged that explained 90% of the overall variance of the system of the 438 patch (each described by 15 variables), while the first factor explained only 52% of the overall variance, the second factor 24%, 8% on the third, and the fourth 6%.

It can take the following list of factors emerged according to the percentage of variance explained: 50% = excellent, 40% = very good, 30% = good, 20% = sufficient, 10% = poor, <= 10% by ignore [15][16].

Table III reports the values of composition of each factor in function of the original variables. The weights of the variables were calculated on each factor (factor loadings).

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

TABLE III. ROTATED FACTOR MATRIX (FACTOR LOADINGS) – WMS 2010 - NUMBER FACTOR 4

Weight of the variables	**F1**	**F2**	**F3**	**F4**
Nzm is correlated with factor 4	0.4374	-0.4517	0.1415	0.5824
Nt is correlated with factor 2	-0.0667	0.9765	0.0660	0.0436
Area is correlated with factor 2	-0.0667	0.9765	0.0660	0.0436
Perimetro is correlated with factor 1	-0.7377	0.2125	-0.1163	-0.5314
DeltaX is correlated with factor 1	-0.9537	0.1984	-0.1589	-0.0500
DeltaY is correlated with factor 1	-0.9272	0.1492	-0.0867	0.1140
AreaIdeale is correlated with factor 1	-0.9469	0.2052	-0.1419	-0.0903
Gx is correlated with factor 3	0.4177	0.0521	0.7174	-0.2025
Gy is correlated with factor 3	0.0733	0.1317	0.9084	0.1544
Jx is correlated with factor 2	-0.2183	0.9404	0.0656	-0.1875
Jy is correlated with factor 2	-0.2189	0.9402	0.0654	-0.1875
Rx is correlated with factor 1	-0.9750	0.1039	-0.1355	-0.0577
Ry is correlated with factor 1	-0.9774	0.0929	-0.1023	-0.0244
AreaRett is correlated with factor 1	-0.9468	0.0773	-0.1085	-0.1710
RapportoAAR is correlated with factor 2	-0.1778	0.5313	-0.3406	0.4860

As can be seen in Figure 8, Figure 9, Figure 10 and Figure 11, there are shown standardized values - greater than the average - calculated on each main component.

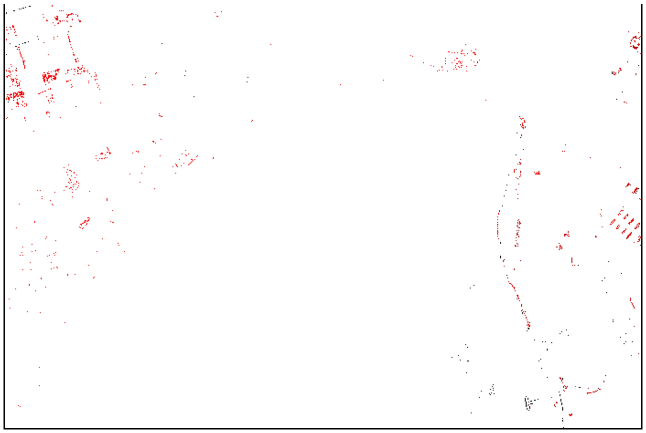


Figure 8. The reconstructed image with the I main component – 2010

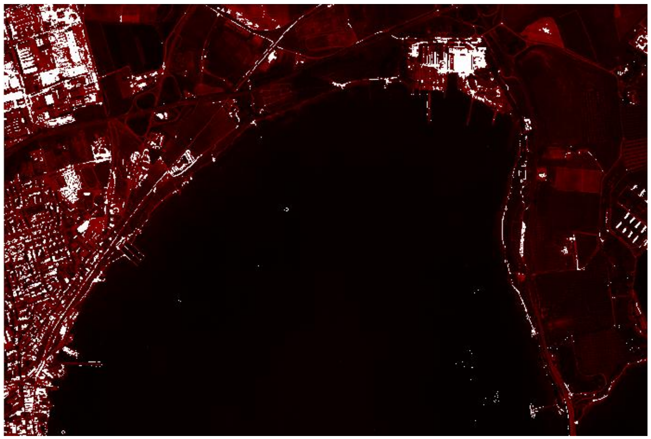


Figure 9. The reconstructed image with the II main component – 2010

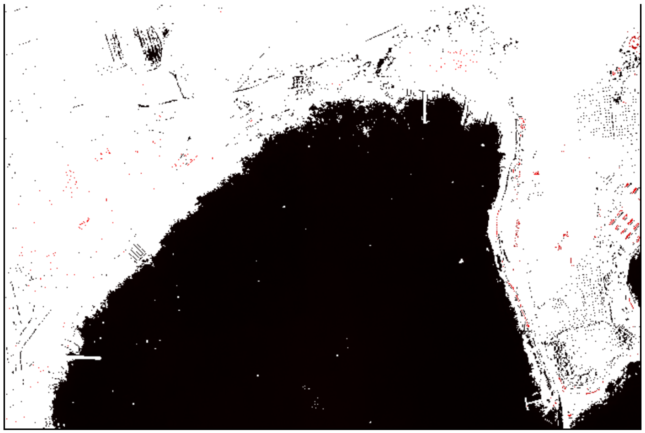


Figure 10. The reconstructed image with the III main component – 2010

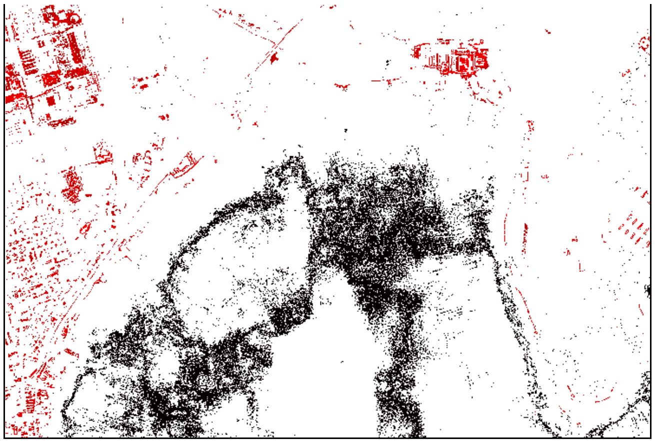


Figure 11. The reconstructed image with the IV main component – 2010

VI. Conclusion and future work

Factor analysis applied, as has been described, the first two images of the Sinus of the Mar Piccolo, respectively in the aerial photo of the Military Geographical Institute that dates back to 1940 and the aerial photo of the Web Map Service (WMS) in 2010, has enabled us to obtain the decomposition of the two analog images into layers that describe numerically the evolutionary peculiarities of the area examined. When considering only the main components characterized by an explanatory contribution to the overall variance of the system more than 10%, it can be seen that the landscape of the first Sinus of the Mar Piccolo of Taranto has suffered fragmentation from 1940 to 2010, a fact that is evidenced by an increase in patches extracted from each of the images taken between 1940 and 2010. In addition, the composition of the loading factors between variables and factors remains generally unchanged between 1940 and 2010, which is meant to signify that there was not a material change in the landscape between 1940 and 2010, with the exception of the arc, where western coastal residential developments grew on significant extensions of territory. It is very interesting to note the comparison of the image reconstructed with the first principal component orthophoto of 1940, (see Figure 3), and the reconstructed image with the fourth main component of the image of the Web Map Service (WMS) 2010 (see Figure 11). Both of these digital images, obtained from the analysis of the main components, seem to indicate what was a significant numerical evolution of the body of water in the first Sinus of the Mar Piccolo between 1940 and 2010. Perhaps it would also be useful to search for physical causes of this evolution.

Acknowledgment

We thank the laboratory of Topography and Photogrammetry of the University of Salento and the Web Map Service (WMS) offered by the geoportal of the Region of Puglia.

References

[1] S. M. Niladri, G. Susmita, and G. Ashish, Fuzzy clustering algorithms incorporating local information for change detection in remotely sensed images Applied Soft Computing, vol. 12, iss. 8, August 2012, pp. 2683-2692.

[2] L. Castellana, A. D’Addabbo, and G. Pasquariello, A composed supervised/unsupervised approach to improve change detection from remote sensing Pattern Recognition Letters, vol. 28, iss. 4, 1 March 2007, pp. 405-413

[3] J. S. Deng , K. Wang , Y. H. Deng & G. J. Qi (2008) PCA based land use change detection and analysis using multitemporal and multisensor satellite data, International Journal of remoote Sensing, 29:16, pp. 4823-4838, DOI: 10.1080/01431160801950162

[4] P. Coppin, I. Jonckheere, K. Nackaerts, and B. Muys, Digital change detection methods in ecosystem monitoring: a review. International Journal of Remote Sensing, 25, pp. 1565-1596, 2004.

[5] J. Byeungwoo, and A. David, Partially supervised classification using weighted unsupervised clustering. IEEE T. Geoscience and Remote Sensing (TGRS) 37(2):1073-1079 1999.

[6] L.G. Browna, Survey of Image Registration Techniques, ACM Computing Surveys, Vol 24, No. 4, December 1992.

[7] K.P. Schwarz, M.A. Chapman, M.E. Cannon, P. Gong, An Integrated INS/GPS Approach to the Georeferencing of Remotely Sensed Data, Photogrammetric Engineering & Remote Sensing, 59(11): 1167-1674, 1993.

[8] H. Hotelling, The generalization of Student’s Ratio. Ann. Math. Statist.,vol. 2, pp. 30-378, 1931.

[9] M. Zevi, The matrix calculation in the method of the components princiapli. Faculty of Architecture, Rome 1977.

[10] F. Ricci, Statistics and Statistical Processing of the Information. Zanichelli, Bologna 1975.

[11] S. Saddocchi, manuale di analisi statistica multivariata F. Angeli, Milano, 1993.

[12] A. Mezzetti, Air pollution and vegetation, Edagricole, 1987.

[13] M.A. Fischler, and R.C. Bolles, Random Sample Consensus: a Paradigm for Mode l Fitting with Applications to Image Analysis and Automated Cartography. Comm. ACM, no. 24, pp. 381-395, 1981.

[14] M. Cramer, D. Stallmann, and N. Halla, High Precision Georeferencing Using GPS/INS and Image Matching, Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Alberta, Canada, June 3-6, pp. 453-462, 1997

[15] G. Zuendorf, N. Kerrouche, K. Herholz, and J.C. Baron, Efficient Principal Component Analysis for multivariate 3D Voxel - based mapping of brain functional imaging data sets as applied to FDG - PET and normal aging, no. 18, pp. 13-21, 2003.

[16] M. Mazzei, and A. L. Palma, Spatial Statistical Models for the Evaluation of the Landscape, Lecture Notes in Computer Science, Computational Science and Its Applications, ICCSA June 2013, pp. 419-432, doi 10.1007/978-3-642-39649-6_30.

Conceptual Modeling for Environmental Niches and Potential Geographic Distributions Using UML GeoProfile

Gerardo José Zárate, Jugurta Lisboa-Filho
Departamento de Informática
Universidade Federal de Viçosa (UFV)
Viçosa, Minas Gerias, Brazil
gerardo.zarate.m@gmail.com, jugurta@ufv.br

Carlos Frankl Sperber
Departamento de Biologia Geral
Universidade Federal de Viçosa (UFV)
Viçosa, Minas Gerias, Brazil
sperber@ufv.br

***Abstract*— An ecological niche is defined by an array of biotic and abiotic requirements that allow organisms to survive and reproduce in a geographic area. Environmental data from a region can be used to predict the potential distribution of a species in a different region. Many formalisms for modeling geospatial information have been developed over the years. The most notable benefit of these formalisms is their focus on a high-level abstraction of reality, leaving unnecessary details behind. This paper presents a conceptual data schema for niches and potential geographic distributions using the UML GeoProfile formalism. The proposed data schema considers the geographic entities and environmental variables involved in the prediction of potential geographic distributions made with ecological niche data.**

Keywords-Geospatial database modeling; Ecological Niches; Potential Geographic Distributions.

I. INTRODUCTION

Conceptual models are formalisms that illustrate entities and relationships between them in a diagram representation. These representations are abstractions of the objects and associations of the real world, leaving unnecessary details out. Database design greatly benefits from conceptual modeling as it focuses on a high-level representation without taking into account implementation details [1][2].

Well-known approaches for modeling databases are the Entity-Relationship (ER) Model introduced by Peter Chen in 1976 [3] and Object-Oriented techniques such as the Object-Oriented Analysis (OOA), Object-Modeling Technique (OMT) and the standard Unified Modeling Language (UML) referred in [2]. These approaches help designers to model databases for almost every human activity.

As Computer Science and technology evolve, there is a necessity to model complex situations in which databases are essential. Databases for Geographic Information Systems (GIS) are a prime example of this. The work of Bédard and Paquette [4] was the first to attempt to include geospatial information in database modeling. They proposed an extension of the ER formalism for modeling spatial data. Since then, many researchers have proposed new formalisms for geospatial data [1][5].

Those formalisms are capable of representing, at the abstract level, geographic features such as roads, buildings or rivers. Moreover, they are also able to represent environmental variables such as temperature or vegetation. The representation and abstraction of geospatial data benefits professionals and scientists in areas, such as Civil Engineering, Agriculture and Ecology, among others.

The ecological niche and potential geographic distributions are fields of study in Ecology that have been of major research interest in the last years [6]. Ecological niches are defined by an array of biotic interactions and abiotic conditions in which a species can survive and reproduce [7]. An environmental niche is constructed only by abiotic conditions [8]. On the other hand, potential geographic distributions refer to areas or regions that have the appropriate set of conditions for a species to live and reproduce. Potential geographic distributions are usually calculated by mathematical algorithms. These algorithms use environment data and occurrences of a species to make predictions [9].

The aim of this paper is to model the entities, relationships and spatial phenomena of environmental niches and potential geographic distributions using a conceptual model for geospatial databases, providing a baseline for the design and implementation of repositories containing ecological niches and potential distribution data.

The rest of this paper is structured as follow: Section II reviews the related work. Section III overviews the basic concepts of ecological niche theory including potential distributions. Section IV offers a summary of geospatial databases formalisms, focusing on UML GeoProfile [2][10]. Section V presents a conceptual data schema for environmental niches and potential geographic distributions and briefly discusses an implementation of the data schema. Finally, Section VI provides some final considerations.

II. RELATED WORK

GIS applications work with geographical features (roads, rivers, buildings) as well as with environmental variables (temperature, humidity, soil). As mentioned in Section I, the aim of this paper is to model niche-based geographic distributions using a formalism for modeling geospatial databases. Previous works have attempted to provide means to model niche and geographic distribution information [9][11][12][13]. This section summarizes prior efforts found in the literature.

ISBN: 978-1-61208-326-1

Although it does not involve conceptual modeling of geospatial databases, the work in [9] emphasizes the importance of databases in GIS applications stressing their storage capabilities. Moreover, it provides a six-step guide for using ecological niche to predict potential geographic distributions. Finally, it describes how environmental variables are handled in GIS applications, highlighting the selection of the appropriate GIS data types.

McIntosh et al. [11] developed a tool that helps ecologists design databases. The focus of their research is to simplify the design process for ecologists with no experience in database theory. They provided previously created templates that help overcome common errors in defining relationships between entities. Models created in their tool can later be exported to a Database Management System (DBMS). The major drawback is the lack of support for geospatial capabilities. Entities cannot be labeled as points, lines, polygons or fields; contrary to conceptual models like those mentioned in Section IV. Even if not directly related to ecological niches or potential distributions, the work in [11] is a valuable effort because it recognizes the importance of databases for ecologists.

Semwayo and Berman [12] presented the guidelines for representing ecological niches in a conceptual model. According to the authors, traditional ER and Object-oriented models fail to represent the granularity of an ecological niche. They propose an ontological engineering approach to model ecological data. Despite the fact that there is no reference to ecological niche theory, the focus of their study is modeling the relationships between humans and their environment. Again, there is no support for geospatial capabilities.

Finally, Keet [13] provides an overview of the principal concepts related to ecological niches and presents an Object-Role Modeling (ORM) diagram of the ecological niche. The proposed ORM diagram includes entities, such as species, fundamental niche, realized niche, hyper-volumes and conditions. The work in [13] is an attempt to model ecological niches based on the concepts first introduced by Grinnell [14] and Hutchinson [15] from a database conceptual standpoint.

Contrary to the described prior work, the data schema proposed in this paper is constructed around data used in niche-based geographic distributions, using a conceptual model with geographic and environmental capabilities. Ultimately, an implementation of the data schema in a DBMS would be capable of storing the necessary geographic and environmental data of ecological niches and potential geographic distributions. Before introducing the proposed schema, it is important to have basic concepts regarding ecological niches, potential geographic distributions and conceptual model for geographic data, which are discussed in Section III and Section IV.

III. Ecological Niche Theory

According to [16], the term ecological niche was first introduced by Joseph Grinnell. Grinnell suggested that a species' niche is defined by its habitat requirements [14]. This means that a niche is determined by all the environmental variables that enable the survival and reproduction of a species.

A similar definition was given by Hutchinson, who introduced the concept of fundamental niche and defined it as an n-dimensional hypervolume determined by species requirements [7][9][15]. Hutchinson's definition is a quantitative approach that gives more clarity to the concept and leaves an open door for the development of mathematical techniques [16].

Although Hutchinson's definition is rather straightforward, an implementation is not a simple task. The amount of dimensions in a hypervolume is potentially infinite. Dimensions such as temperature and soil characteristics can be easy to collect, while other variables like the diet of an organism are, in some cases, not accessible. Additionally, certain dimensions can be irrelevant to determine the fundamental niche [7][15].

The dimensions of the hypervolume can be classified as conditions and resources. Resources are consumed or used, which might lead to competition between organisms of the same or different species. Differently, conditions are environmental (abiotic) variables, such as temperature, precipitation and terrain aspect, among others [8].

Depending on the dimensions considered, ecological niches can be classified as Grinnellian or Eltonian. Grinnellian niches (also referred as environmental niches) consider only environmental variables, which are, in most cases, considered scenopoetic, i.e., not affected by organisms. On the other hand, Eltonian niches focus on resources and relationships between organisms. The concept of n-dimensional hypervolume can be applied to both Grinnellian and Eltonian niches [8]. This paper, considers only environmental niches, as their data sets are becoming more available and data sets for Eltonian niches are still difficult to obtain [8]. Furthermore, data from environmental niches are more related to predictions of geographic distributions, which are also in the scope of this paper [9].

Others exploited concepts related to ecological niches are the realized niche and the geographic distribution of species. Hutchinson defined the realized niche as a subset of the fundamental niche restricted by species' biotic interactions [9][16]. According to Soberón, the realized niche occurs in the overlapping area between the geographic region with appropriate abiotic factors and the region in which there is a suitable combination of interaction between species [17]. The actual geographic distribution of a species would be the region that has the appropriate range of abiotic and biotic conditions, as well as being accessible to organisms [17][18]. The BAM Diagram (called BAM due to the labels in each circle of the diagram) [17] exhibited in Fig. 1 offers a graphic explanation of the concepts defined earlier. The circle A represents the area with the appropriate abiotic conditions (geographical expression of the fundamental niche). The circle B is the area with suitable combination of interacting species. The intersection of A and B denotes the geographical extent of the realized niche. Circle M holds the regions accessible to the species. Finally, the overlapping region of A, B and M represents the geographic distribution.

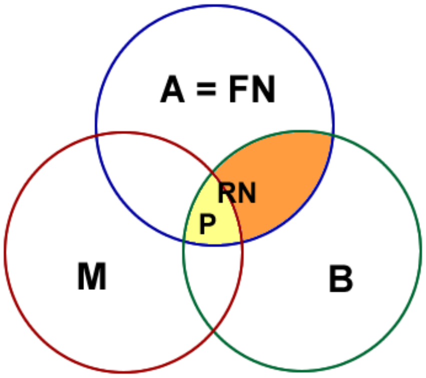


Figure 1. BAM Diagram for representing the fundamental niche [17]

Additionally to the actual geographic distribution, potential geographic distributions are regions with suitable conditions for species to survive, i.e., the geographical extent of the fundamental niche [9][19]. Usually, data from species distribution (occurrences and environmental variables) are used in mathematical algorithms to predict potential geographic distributions [9][19][20][21]. The inputs for these algorithms are a set of occurrence data and environmental variables for both occupied and evaluated area. Outputs, on the other hand, are either regions with suitable conditions in which species are present (the intersection P of the three regions of the BAM diagram), or regions with suitable conditions where organisms are not present (areas representing the fundamental niche A minus P in the BAM diagram) [17].

As mentioned in [6] and [22], since the 1990s, the methodologies based on Ecological Niche Modeling have increased significantly. There are several uses for niche related concepts in the literature including climate change projections, potential geographic distributions, species invasion projections, niche characterization, niche diversification, niche construction and habitat-suitability, among others [6][8][21][23][24][25].

IV. CONCEPTUAL MODELS FOR GEOSPATIAL DATABASES

Over time, computational systems have become more robust and sophisticated; hence, there is a necessity to handle complex data such as geospatial information. One of the major elements of a GIS is a database in which information is stored. Modern DBMS software, such as Oracle and PostgreSQL, have capabilities to manage geospatial data and provide additional benefits like security, redundancy or user control access.

Database designing has three basic stages [26]: conceptual, logical and physical. The conceptual stage produces data schemas that represent a high-level abstraction of entities and relationships between them. The major benefit of using conceptual models is their independence of implementation details, which is the reason of their usage in Computer Science fields such as Databases. Notable conceptual models used in database modeling are the ER Model, OOA, OMT and the UML [2][3].

The work in [4] was the first attempt to create a conceptual model (formalism) dedicated to model geospatial databases from a conceptual standpoint. Bédard and Paquette proposed a geospatial extension of the ER formalism. Thenceforth, many researchers and professionals have proposed new methods or extended previous ones. Conceptual models for geospatial databases assist in the process of modeling geographical features as they are modeled as perceived by humans [27]. Moreover, the studies in [28] and [29] state that geospatial formalisms allow reduction in the number of entities and relationships without losing semantics.

The studies in [1] and [5] present a timeline of the major geospatial formalisms and list their principal characteristics. According to Pinet [1], there are seven major goals shared by formalisms dedicated to model geospatial data:

1) Representing basic geospatial objects such as points, lines, polygons, multiple points, multiple lines or multiple surfaces.
2) Modeling geospatial relationships between objects. Examples of relationships are adjacency, overlap and disjoint.
3) Description of the evolution of objects over time.
4) Modeling objects that might have multiple representations depending of the geographical scale.
5) Description of objects with uncertain boundaries or positions, for instance floods or areas of pollution.
6) Representation of continuous geospatial data that can be measured in any location of the study area.
7) Modeling structured networks.

Usually, formalisms use pictograms to improve readability and to simplify the model [5]. A pictogram is a graphic symbol that resembles the real object that is being modeled. Fig. 2 shows the pictograms used in UML GeoProfile [10]. Notice that the pictograms cover most of the goals proposed in [1].

Comparing the various formalisms specialized in geospatial data is not in the scope of this paper. For a comparison and overview of different formalisms, refer to [1][2][5].

A. UML GeoProfile Overview

UML GeoProfile is an UML profile specifically designed as a formalism for modeling geospatial databases in a conceptual level. As noticed before, a conceptual model represents an abstraction of reality and does not involve implementation details. Being an UML extension, UML GeoProfile allows the use of classes, associations, packages and constraints, among other UML features [2]. Additionally, UML GeoProfile can be implemented in any Computer Aided Software Engineering (CASE) tool with UML profiles support.

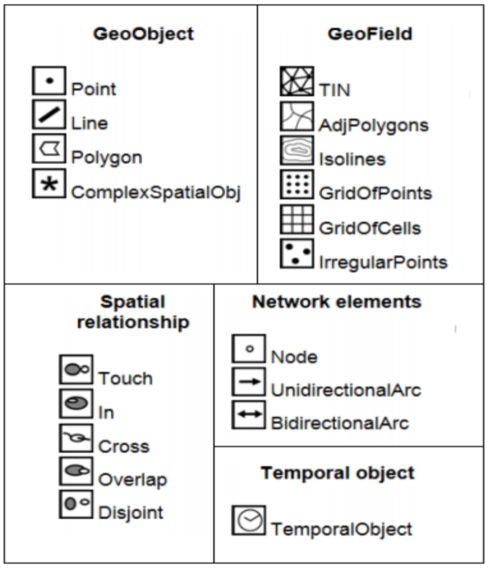


Figure 2. Pictograms used in UML GeoProfile [10].

The principal motivation behind UML GeoProfile was the standardization of previous models. To accomplish this, UML GeoProfile took the best offerings from different models and brought them together. As other formalisms, UML GeoProfile takes advantage of pictograms to simplify the model. In UML GeoProfile, pictograms are modeled as stereotypes. A UML stereotype allows designers to extend the terminology of UML in order to create new constructors [2][10]. Furthermore, UML GeoProfile takes advantage of UML packages to divide schemas in geospatial themes, e.g., vegetation, relief or hydrography. This characterizes related entities and provides better organization.

UML GeoProfile follows the international standards for Geographic Information of the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC) [10], which reduce inconsistencies between *de jure* and *de facto* standards [30]. Additionally, UML GeoProfile adopts a Model-Driven Architecture (MDA) approach. In MDA, models are first built in a Computation Independent Model (CIM); CIM models are later transformed to a Platform Independent Model (PIM). The third stage of the process is the Platform Specific Model (PSM), which is later converted to implementation code [10]. Further information on stereotypes, international standards and MDA can be found in [2][10][30].

V. Representing Environmental Niches and Potential distributions Using UML GeoProfile

This section describes how to model environmental niches and potential geographic distributions using UML GeoProfile as an MDA's Computational Independent Model (CIM). First, we illustrate the representation of individual entities, and, then, we propose an approach to represent environmental niches and potential geographic distributions in three packages that form a single database schema.

A. Basic representations

As stated in Section IV, UML GeoProfile uses stereotypes to represent geospatial entities (classes in UML). For instance, the *Point* stereotype is used to represent trees or occurrence data while the *Polygon* stereotype represents geographic areas such as cities or forests. Moreover, UML GeoProfile provides stereotypes to represent continuous data such as humidity and temperature. Fig. 3 exhibits a representation of occurrence data of species (a), the occupied area in which organisms live (b) and temperature (c) employing UML GeoProfile.

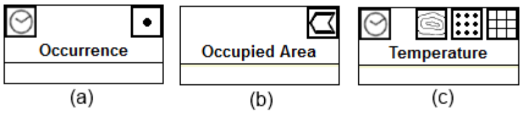


Figure 3. Representations of occurrence data of a species (a), occupied area (b) and multiple representations of temperature (c).

Additionally, UML GeoProfile supports multiple representations for geospatial entities. Depending on how data was initially collected, environmental variables can be represented in a diversity of GIS types. Fig. 3 (c) shows the representation of temperature displayed as Isolines, Grid of Points and Grid of Cells.

B. Modeling Environmental Niches

Predictive algorithms such as the Genetic Algorithm for Rule-Set Prediction (GARP) [31] work with a set of occurrence data and an array of environmental variables of a given area to predict geographic distributions. A range is defined for each variable (minimum and maximum values) to construct an n-dimensional hypervolume that restricts the conditions in which organisms can survive.

Being a generic modeling approach, it is necessary to build a schema that supports data from different species and multiple regions occupied by organisms of the same species. Each region has its own hypervolume defined by an array of environmental variables. Furthermore, the amount of variables can also vary from region to region.

Fig. 4 presents the proposed conceptual data schema for environmental niche data. Notice that the entities and their relationships are based on the literature for ecological niches and potential geographic distributions referenced in Section III. An occupied area (modeled with the Polygon stereotype) has one or multiple occurrences of a species. The relationship between the Occupied Area and Occurrence classes is modeled as a spatial relationship *In*, indicating that a species occurrence is inside a region. The term “occurrence” is preferred over “organism” because the schema does not consider particular characteristics of organisms such as weight or age. The organism's location (covered by the Point stereotype) is the most important piece of information. In addition, the schema does not take into consideration organisms' movement. For that reason, an occurrence is related to not more than one occupied area. However, an organism can be identified in two or more areas over a period of time. Although an occurrence can represent the same organism, it is still related to at most one region in a particular moment. To solve the relationship's lack of congruence, the Temporal Object stereotype is also assigned to the Occurrence class. This allows an organism to be related to other regions in a different moment in time.

Multiple environmental variables can be considered in an occupied area. In addition, a type of environmental variable can be analyzed in multiple regions as well. Here, the hypervolume is defined by the multiple instances of the Niche Axis (hypervolume dimension) association class, which cannot exist without the association between occupied areas and environmental variables.

An association class is defined for each association between the two classes, indicating the units (ratio, degrees, inches) and the minimum and maximum values of a particular variable in a specific area. This approach presents a limitation: it is incapable of modeling relationships between dimensions of the hypervolume, e.g., if the temperature is higher than 30◦C, then the humidity must be between 90% and 99% [14]. These relationships depend entirely on the environmental variables and their variation. Consequently, it is difficult to predict and model them. Analytical tools or algorithms handle the relationships as rule sets used to predict geographic distributions [9][21].

Notice that in Fig. 4, a GeoField stereotype is not assigned to the Environmental Variable class as it was previously suggested. The reason behind this is that GeoObject classes (points, lines, polygons) and GeoField classes belong to different views of the reality and usually there are not topological relationships between classes of two different views [10]. That said, the lack of stereotype for the class is a non-issue. In order to construct a hypervolume it is only necessary to know the variable type and its range. Nevertheless, it is also important to provide a manner of including in the model the field from which the hypervolume data were extracted.

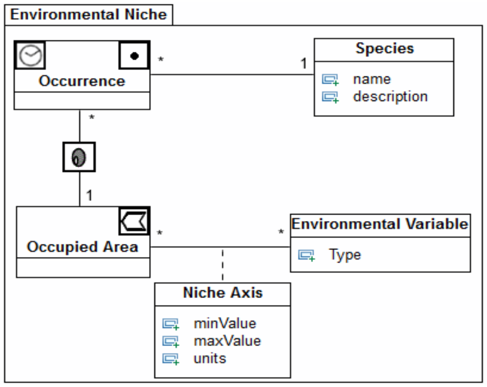


Figure 4. Representation of environmental niches

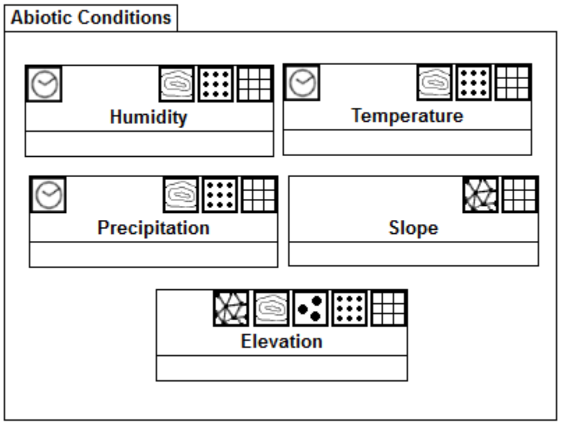


Figure 5. Possible environmental variables of an n-dimensional hypervolume. Multiple representations allow the use of different types of data sources.

As mentioned before, the amount of dimensions in a hypervolume is potentially infinite. Hence, the final model strictly depends on the study case. Fig. 5 provides an example of the possible representation of the environmental dimensions (abiotic conditions) of a hypervolume. Notice the presence of the Temporal Object stereotype in some classes, meaning that certain abiotic conditions can vary over time, e.g., the monthly average temperature of a region.

C. *Representing Potential Geographic Distributions*

Predictive algorithms and tools operate with occurrence data and environmental variables to produce potential geographic distributions of a species (regions where organisms can live or survive) usually in the form of a grid of cells [9][19]. Fig. 6 shows the classes related to the potential distribution.

The Evaluated Region class represents the boundaries of the area in which the distribution is projected; this is relevant to model because projections are usually done from a defined area to another. For example, the research in [9] used niche data of a pathogen from the United States (occupied area) and predicted distributions for Mexico (evaluated region); similar researchers are found in [19] and [21]. Evidently, environmental data from the evaluated region are also needed. These data are modeled in the same manner as the environmental dimensions of the niche hypervolume (refer to Fig. 5).

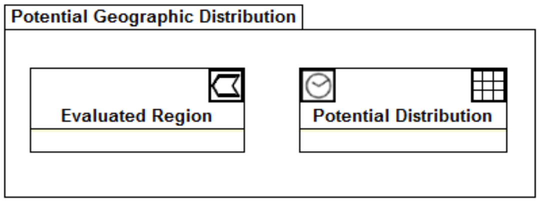


Figure 6. Potential Geographic Distribution. The evaluated region is modeled as polygon and the distribution as a grid of cells.

Notice that the Evaluated Region and Potential Distribution classes are not associated because they belong to different views. Additionally, there is no relationship between a species (or its organisms) and the evaluated regions. Even if organisms occupy the latter, there is no evidence of a topological relationship.

Finally, it is inevitable to acknowledge the necessity to link the field view classes (both abiotic conditions and Potential Distribution classes) to their corresponding region. This can be done through metadata that describes details such as coverage area or how and when data were obtained.

D. Implementation of the data schema

We implemented the conceptual data schema in PostgreSQL using the PostGIS geospatial extension and its *geometry* and *raster* data types to store geographic and environmental data. Non-geospatial entities were implemented using basic data types provided by the DBMS. To employ the data schema, first we took advantage of the software openModeller [32] to create the potential geographic distribution and ecological niche model of sample data (occurrences and environmental data) provided with openModeller.

The results generated by one of the algorithms included in openModeller were later stored in the data schema using basic SQL statements and tools designed to load geospatial information into a PostgreSQL database. QuantumGIS and other GIS software with geographic analytical capabilities can be used to retrieve the information stored in the database (information can be filtered by species, area of interest, among others). This provides the benefit of having data for multiple species stored in a single place instead of different files. Furthermore, our approach can exploit all the advantages of a DBMS.

VI. Concluding Remarks

This paper presented a conceptual data schema for environmental niches and potential geographic distribution of species. The complete schema consists of the components exhibited in Figures 4, 5 and 6. The major limitations of this approach are the lack of support for relationships between dimensions of the niche's hypervolume and the inability to model classification values such as vegetation type. Both limitations are handled by predictive algorithms in a form of rule sets generated from the abiotic layers.

The geospatial and temporal phenomena of the schema are modeled using UML GeoProfile stereotypes. UML GeoProfile was preferred over other formalisms for its capacity to model both object and field phenomena, as well as for the implementation of international standards, and MDA adoption. The proposed conceptual data schema represents no more than the CIM stage of the MDA. Future work includes implementation of the remaining MDA stages and development of a study case with real data.

Acknowledgment

This project was partially financed by CAPES, OAS, CNPq and FAPEMIG.

References

[1] F. Pinet, "Entity-relationship and object-oriented formalisms for modeling spatial environmental data." Environmental Modelling & Software 33, 2012, pp. 80-91.

[2] J. Lisboa-Filho, G. Sampaio, F. Nalon and K. Borges. "A UML profile for conceptual modeling in GIS domain". Proceedings of the International Workshop on Domain Engineering at CAiSE. Hammamet, Tunisia, 2010, pp. 18-31.

[3] P. Chen, "The entity-relationship model—toward a unified view of data." ACM Transactions on Database Systems (TODS) 1.1, 1976, pp. 9-36.

[4] Y. Bédard and F. Paquette, "Extending Entity/Relationship Formalism for Spatial Information Systems". AUTO-CARTO 9, Baltimore, 1989, pp. 818-827.

[5] A. Miralles, F. Pinet, and Y. Bédard. "Describing spatio-temporal phenomena for environmental system development: An overview of today's needs and solutions." International Journal of Agricultural and Environmental Information Systems 1.2, 2010, pp. 68-84.

[6] A. Peterson and J. Soberon. "Species distribution modeling and ecological niche modeling: getting the concepts right." Natureza & Conservação 10, no. 2, 2012, pp. 102-107.

[7] J. Polechová and D. Storch; "Ecological Niche" Encyclopedia of Ecology, Vol. 2, 2008, pp. 1088-1097, eds. Sven Erik Jørgensen and Brian D. Fath, Oxford: Elsevier.

[8] J. Soberón, "Grinnellian and Eltonian niches and geographic distributions of species." Ecology letters 10.12, 2007, pp. 1115-1123.

[9] J. Blackburn, "Integrating geographic information systems and ecological niche modeling into disease ecology: a case study of Bacillus anthracis in the United States and Mexico." Emerging and Endemic Pathogens, Springer Netherlands, 2010, pp 59-88.

[10] J. Lisboa-Filho, F. Nalon, D. Peixoto, G. Sampaio, and K. Borges, "Domain and Model Driven Geographic Database Design". In Domain Engineering: Product Lines, Languages, and Conceptual Models, 2013, pp. 375-399.

[11] A. McIntosh, J. Cushing, N. Nadkarmi and l. Zeman, "Database design for ecologists: Composing core entities with observations." Ecological informatics 2.3, 2007, pp. 224-236.

[12] D. Semwayo and S. Berman. "Representing ecological niches in a conceptual model." Conceptual Modeling for Adv. App. Domains. Springer Berlin Heidelberg, 2004, pp. 31-42.

[13] M. Keet, "Representations of the ecological niche." WSPI 2006: Contributions to the Third International Workshop on Philosophy and Informatics vol. 14, 2006, pp. 75-88.

[14] J. Grinnell, "The niche-relationships of the California Thrasher." The Auk34, no. 4, 1917, 427-433.

[15] G. Hutchinson, "Concluding remarks". Cold Spring Harbour Symposium on Quantitative Biology 22, 1957, pp. 415–427.

[16] J. Chase and M. Leibold, "Ecological niches: linking classical and contemporary approaches". University of Chicago Press, 2003.

[17] J. Soberon, "Interpretation of models of fundamental ecological niches and species' distributional areas." Biodiversity Informatics vol. 2, 2005, pp. 1-10.

[18] N. Sillero, "What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods." Ecological Modelling 222.8, 2011, pp. 1343-1346.

[19] D. Ward, "Modelling the potential geographic distribution of invasive ant species in New Zealand." Biological Invasions 9.6, 2007, pp. 723-735.

[20] M. De Meyer, et al. "Ecological niche and potential geographic distribution of the invasive fruit fly Bactrocera invadens (Diptera, Tephritidae). Bulletin of Entomological Research 100.1, 2010, pp. 35-48.

[21] A. Peterson, "Predicting the geography of species' invasions via ecological niche modeling." The quarterly review of biology 78.4, 2003, pp. 419-433.

[22] H. Owens, et al. "Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas." Ecological Modelling 263, 2013, pp. 10-18.

[23] K. Laland and N. Boogert, "Niche construction, co-evolution and biodiversity." Ecological Economics 69.4, 2010, pp. 731-736.

[24] A. Jiménez-Valverde, A. Peterson, J. Soberón, J. Overton, P. Aragón and J. Lobo. "Use of niche models in invasive species risk assessments." Biological Invasions 13.12, 2011, pp. 2785-2797.

[25] A. Hirzel, J. Hausser, D. Chessel, and N. Perrin, "Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data?" Ecology 83.7, 2002, pp. 2027-2036.

[26] R. Elmasri, S. Navathe, “Fundamentals of Database Systems” (6th Ed.), Addison Wesley, Boston, MA, 2010.

[27] K. Borges, C. Davis, and A. Laender, "OMT-G: an object-oriented data model for geographic applications." GeoInformatica 5.3, 2001, pp. 221-260.

[28] C. Parent, S. Spaccapietra, E. Zimanyi, P. Donini, C. Plazanet and C. Vangenot. "Modeling spatial data in the MADS conceptual model." Int. Symp. on Spatial Data Handling, Vancouver, 1998, pp. 138-150.

[29] Y. Bédard, C. Caron, Z. Maamar, B. Moulin and D. Vallière, “Adapting data models for the design of spatio-temporal databases”. Computers, Environment and Urban Systems, Vol. 20, Issue 1, 1996, pp. 19-41.

[30] J. Brodeur and T. Badard, "Modeling with iso 191xx standards" Encyclopedia of GIS, 2008, pp. 705-716. 07, IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.

[31] D. Stockwell, "The GARP modelling system: problems and solutions to automated spatial prediction." Int. Journal of Geographical Information Science 13.2, 1999, pp. 143-158.

[32] M.E. de Souza Muñoz, et al. "openModeller: a generic approach to species' potential distribution modelling." GeoInformatica 15, no. 1, 2011, pp.111-135.

Spatial Visibility Clustering Analysis in Urban Environments Based on Pedestrians' Mobility Datasets

Oren Gal, Yerach Doytsher
Mapping and Geo-information Engineering
Technion - Israel Institute of Technology
Haifa, Israel
e-mail: {orengal,doytsher}@technion.ac.il

Abstract—**In this paper, we propose a Spatial Visibility Clustering (SVC) method estimating the number of clusters (groups) k, based on 3D visible volumes analysis in urban environments. We extend our previous work and propose fast and exact 3D visible volumes analysis in urban scenes based on an analytic solution. We test and analyze the SVC method by using real records of pedestrians' mobility datasets from the city of Melbourne and by setting control points for efficient monitoring and control using a K-means clustering algorithm. By testing large databases, we also propose time zones division for optimal control points.**

Keywords-Visibility; 3D; Urban environment; Spatial analysis.

I. INTRODUCTION AND RELATED WORK

'Clustering methods' refers to the division of data sets into groups, each containing similar objects. Data modeling is extensively studied in statistics, mathematics and machine learning [1]. Most of the common clustering methods can be divided into hierarchical and partitioning methods.

Partitioning algorithms determine the clusters directly, such as the well-known K-Means method, where by a hierarchical mechanism, builds the clusters gradually.

Clustering methods of 2D spatial data (such as GIS database) were also studied, defining data proximity by using, inter alia, a Delaunay diagram. These methods focused on performances and low complexity, by keeping K-nearest neighbors using a connectivity graph where clusters become connected components [6].

Our research contributes to the spatial data clustering field, where, as far as we know, visibility analysis has become a leading factor for the first time. The SVC method, while mining the real pedestrians' mobility datasets, enables by a visibility analysis to set the number of clusters.

The efficient computation of visible surfaces and volumes in 3D environments is not a trivial task. Accurate visibility computation in 3D environments is a very complicated task demanding a high computational effort, which could hardly have been done in a very short time using traditional well-known visibility methods [15].

Most of these works have focused on approximate visibility computation, enabling fast results using interpolations of visibility values between points, calculating point visibility with the Line of Sight (LOS) method [4,5]. Lately, fast and accurate visibility analysis computation in 3D environments has been presented [7,8,9,10].

In this paper, we introduce a unified method for estimating the number of clusters using 3D visible volumes analysis, SVC. Based on our previous work, we use a fast and efficient analytic solution, setting visibility boundaries of visible surfaces from the viewpoint. We extend our solution to 3D volumes, computing 3D visible volumes. By using *F*-criteria, we set the optimal number of clusters from the visibility aspect.

We demonstrate our method using real datasets from the city of Melbourne's 24-hours pedestrians monitoring system, localizing control points at each hour during the day, using a K-means algorithm with SVC output, i.e., number of clusters *k*. We analyze pedestrians' mobility behavior and suggest dividing the day into four time zones, based on our datasets and setting optimal control points during these time zones.

In Section II, we first introduce the SVC method, and the extended visible volumes analysis. In Section III, we present the SVC simulation using the city of Melbourne's datasets. Eventually, we present our approach by dividing a day's hours into four time zones and setting optimal control points.

II. SPATIAL VISIBILITY CLUSTERING (SVC) METHOD

We present, for the first time as far as we know, a unified spatial analysis defining the number of clusters in a data set based on analytic visibility analysis, SVC. The output of our method can be efficiently used by common clustering methods (e.g., K-means or hierarchical). The number of clusters in dense environments can be used for civil and security applications in urban environments, based on 3D visibility analysis from points of view.

For the last twenty years, many methods were proposed in order to estimate the number of clusters in data sets [2,3,11,18]. As previously mentioned [11], the approaches can be divided into global and local methods.

First, we introduce the main steps of our method and formulate the problem of estimating the number of clusters and the proposed volumes visibility analysis in 3D. Later, we present the analysis of the number of clusters using the SVC method, based on real pedestrians' mobility data sets. Finally,

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

we examine a unique division of a twenty four-hour day into four different time zones in Melbourne [13], for control points based on pedestrians' mobility datasets in a number of points of interest, presented in Figure 1.

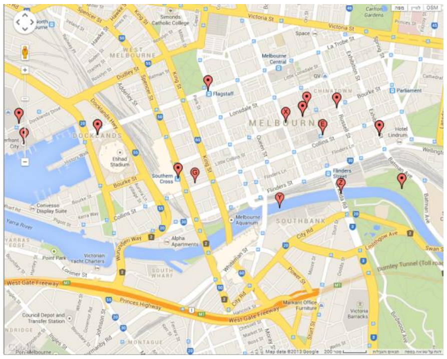


Figure 1. Melbourne Sensors Location for Monitoring Pedestrians' Mobility Data

A. Spatial Visibility Clustering - Main Stages

Our data set $\{X_{ij}\}$, *i=1,2,..,n, j=1,2,..,p,* consists of *p* features measured on *n* independent observation, where observation points marked with blue circles are illustrated in Figure 2. We clustered the data into *k* clusters, $C_1, C_2, .. C_k$. For cluster r, denote as C_r with n_r observations:

$$V_r = \sum_{i\in C_r}\sum_{j\in C_r}\left\|V(x_i)-V(x_j)\right\| \qquad (1)$$

$$V_r = \sum_{i\in C_r}\left\|V(x_i)-V(\bar{x})\right\| \qquad T_k = \sum_{r=1}^{k}\frac{1}{S}V_r$$

where $V(x)$ denotes the visible volumes from a viewpoint *x* bounded inside the total volume *S*, V_r is the sum of the absolute visibility differences of all viewpoints from their cluster visibility mean, and the normalized visible volumes T_k for all clusters *r=1..k*, called ***dispersion***.

Similarly to many other methods estimating the number of clusters [11,17], we define reference data sets distributed uniformly inside bounding volume *S*. We define our reference data sets with the same size of the original data set *X*, and calculate the dispersion of these datasets, T_k^*.

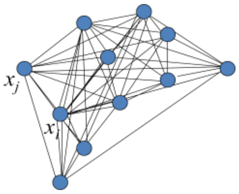


Figure 2. Pedestrians' location architecture based on monitoring datasets, observation points marked with blue circles

Based on *F* statistic, datasets are analyzed, where adding another cluster does not give a better modeling of the data, also known as *F-test* criteria. By setting a group's visibility variance, the number of clusters can be estimated efficiently:

$$SVC_n(k) = T_k^* - T_k \qquad (2)$$

Fast and efficient visibility volume computation from a specific viewpoint, bounded in volume *S*, is presented in the next subsection.

SVC steps can be summarized as follows:

1. Calculate the sum of absolute visibility differences of all points from their cluster visibility mean. Normalize this sum for all possible clusters T_k, also called dispersion.
2. Generate a set of reference datasets, simulated by a uniform distribution model inside bounding volume *S*.
3. Calculate the dispersion of each of these reference datasets, and calculate their mean visibility values.
4. Define SVC for each possible number of clusters as: Expected dispersion of reference datasets - Dispersion of original dataset.

Originally, *F* statistic was used to test the significance of the reduction in the sum of squares as we increase the number of clusters [11]. In general, when the number of clusters increases, the in-cluster decay first declines rapidly. From a certain *k*, dividing a dataset into *k+1* clusters decreases the value of *F-test* function which depends on *k*.

Approximated *F-test* function: Assuming that T_k is the partition of n instances into *k* clusters, and T_{k+1} is obtained from T_k splitting one of the clusters, then the overall mean ratio can be approximated as:

$$F_k = \frac{SVC_n}{SVC_{n+1}} \qquad (3)$$

We adapted aspects of previous F statistic theory for visibility analysis. More detailed F statistic analysis can be found in [11].

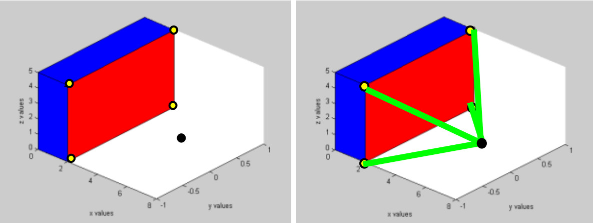
The spatial meaning of this mathematical clustering formulation can be simplified as a group of viewpoints with minimal difference to the average visible volume in the same bounding box.

B. Analytic 3D Visible Volumes Analysis

In this section, we present fast 3D visible volumes analysis in urban environments, based on an analytic solution which plays a major role in our proposed method of estimating the number of clusters. We extend our previous work [7] for surfaces visibility analysis, and present an efficient solution for visible volumes analysis in 3D.

We analyze each building, computing visible surfaces and defining visible pyramids using analytic computation for visibility boundaries [7]. For each object we define Visible Boundary Points (VBP) and Visible Pyramid (VP).

A simple case demonstrating analytic solution from a visibility point to a building can be seen in Figure 3(a). The visibility point is marked in black, the visible parts colored in red, and the invisible parts colored in blue where VBP marked with yellow circles.



(a) (b)

Figure 3. (a) Visibility Volume Computed with the Analytic Solution. (b) Visible Pyramid from a Viewpoint (marked as a Black Dot) to VBP of a Specific Surface

In this section, we introduce our concept for visible volumes inside bounding volume by decreasing visible pyramids and projected pyramids to the bounding volume boundary. First, we define the relevant pyramids and volumes.

The Visible Pyramid (VP): we define $VP_i^{j=1..Nsurf}(x_0, y_0, z_0)$ of the object i as a 3D pyramid generated by connecting VBP of specific surface j to a viewpoint $V(x_0, y_0, z_0)$.

In the case of a box, the maximum number of N_{surf} for a single object is three. VP boundary, colored with green arrows, can be seen in Figure 3(b).

For each VP, we calculate Projected Visible Pyramid (PVP), projecting VBP to the boundaries of the bounding volume S.

Projected Visible Pyramid (PVP) - we define $\mathrm{PVP}_i^{j..N_{surf}}(x_0, y_0, z_0)$ of the object i as 3D projected points to the bounding volume S, VBP of specific surface j trough viewpoint $V(x_0, y_0, z_0)$. VVP boundary, colored with purple arrows, can be seen in Figure 4.

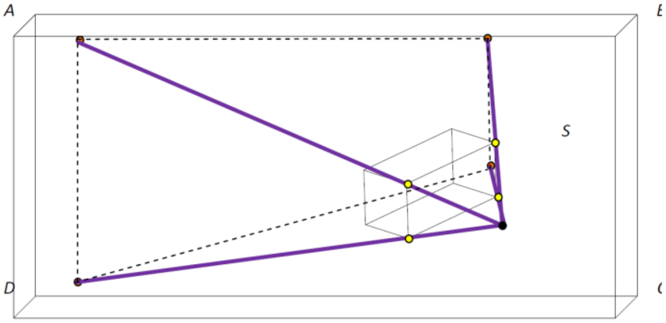


Figure 4. Invisible Projected Visible Pyramid Boundaries colored with purple arrows from a Viewpoint (marked as a Black Dot) to the boundary surface ABCD of Bounding Volume S

The 3D Visible Volumes inside bounding volume S, $\mathrm{VV_S}$, computed as the total bounding volume S, $\mathrm{V_S}$, minus the Invisible Volumes $\mathrm{IV_S}$. In a case of no overlap between buildings, $\mathrm{IV_S}$ is computed by decreasing the visible volume from the projected visible volume, $\sum_{i=1}^{N_{obj}}\sum_{j=1}^{N_{surf}}(\mathrm{V}(\mathrm{PVP}_i^j) - \mathrm{V}(\mathrm{VP}_i^j))$.

$$\mathrm{VV_S} = \mathrm{V_S} - \sum_{i=1}^{N_{obj}}\sum_{j=1}^{N_{surf}} \mathrm{IV}_{S_i}^j \qquad (4)$$

$$\mathrm{VV_S} = \mathrm{V_S} - \sum_{i=1}^{N_{obj}}\sum_{j=1}^{N_{surf}} (\mathrm{V}(\mathrm{PVP}_i^j) - \mathrm{V}(\mathrm{VP}_i^j))$$

By decreasing the invisible volumes from the total bounding volume, only the visible volumes are computed, as seen in Figure 5. Volumes of VPV and VP can be simply computed based on a simple pyramid volume geometric formula.

In a case of two buildings without overlapping, $\mathrm{IV_S}$ computed for each building, as presented above, as can be seen in Figure 6.

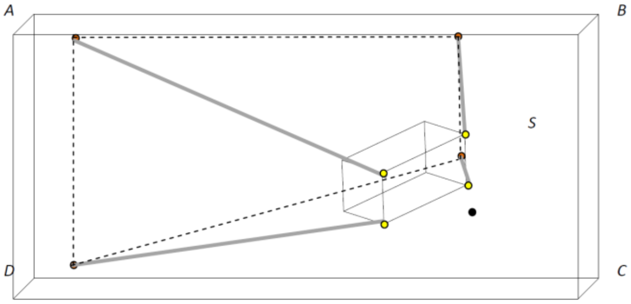


Figure 5. Invisible Volume $\mathrm{V}(\mathrm{PVP}_i^j) - \mathrm{V}(\mathrm{VP}_i^j)$ Colored in Gray Arrows. Decreasing Projected Visible Pyramid boundary surface ABCD of Bounding Volume S from Visible Pyramid

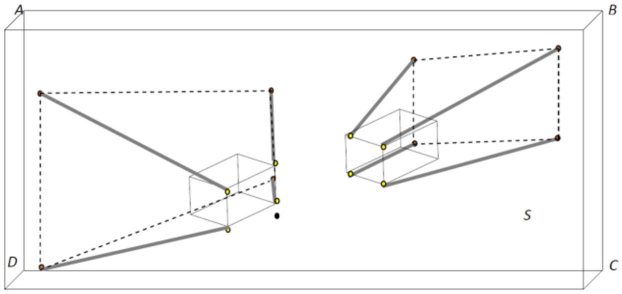


Figure 6. Invisible Volume $\mathrm{V}(\mathrm{PVP}_i^j) - \mathrm{V}(\mathrm{VP}_i^j)$ Colored in Gray Arrows. Decreasing Projected Visible Pyramid boundary surface ABCD of Bounding Volume S from Visible Pyramid

Considering two buildings with overlap between object's Visible Pyramids, as seen in Figure 7(a). In Figure 7(b), VP_1^1 boundary is colored by green lines, VP_2^1 boundary is colored by purple lines and the hidden and Invisible Surface between visible pyramids $\mathrm{IS}_{\mathrm{VP}_1^i}^{\mathrm{VP}_2^i}$ is colored in white.

Invisible Hidden Volume (IHV) - We define Invisible Hidden Volume (*IHV*), as the *Invisible Surface (IS)* between visible pyramids projected to bounding box *S*.

For example, IHV in Figure 7(c) is the projection of the invisible surface between visible pyramids colored in white, projected to the boundary plane of bounding box *S*.

In the case of overlapping buildings, by computing invisible volumes IV_S, we decrease IHV twice between the overlapped objects, as can be seen in Figure 7(c), *IHV* boundary points denoted as $\{A_{11}, .., A_{18}\}$. The same scene is presented in Figure 8, where Invisible Volume $V(PVP_i^j) - V(VP_i^j)$ is colored in purple and green arrows for each building.

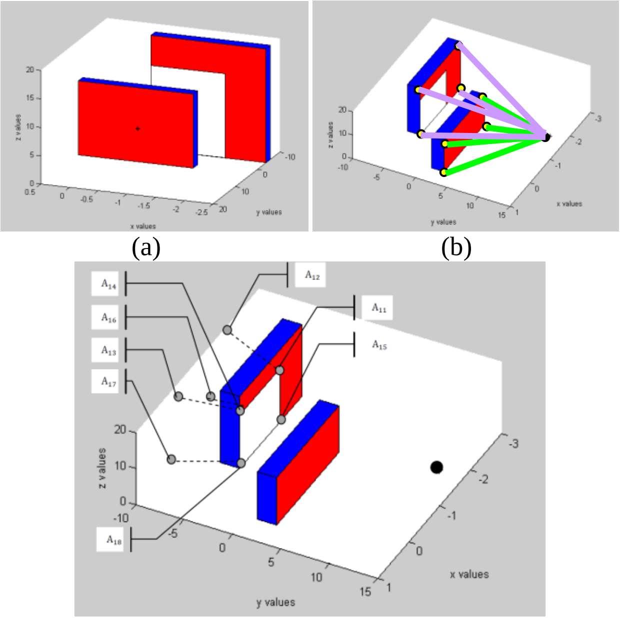


Figure 7. (a) Computing Hidden Surfaces between Buildings , *VP₂¹* Base Plane, $IS_{VP_1^i}^{VP_2^i}$ (b) The Two Buildings - *VP₁¹* in green and *VP₂¹* in Purple (from the Viewpoint) and $IS_{VP_1^i}^{VP_2^i}$ in White (c) IHV boundary points colored with gray circles denoted

The *PVP* of the object close to the viewpoint is marked in black, colored with pink circles denoted as boundary set points $\{B_{11}, .., B_{18}\}$ and the far object's *PVP* is colored with orange circles, denoted as boundary set points $\{C_{11}, .., C_{18}\}$. It can be seen that *IHV* is included in each of these invisible volumes, where $\{A_{11}, .., A_{18}\} \in \{B_{11}, .., B_{18}\}$ and $\{A_{11}, .., A_{18}\} \in \{C_{11}, .., C_{18}\}$.

Therefore, we add *IHV* between each overlapping pair of objects to the total visible volume. In the case of overlapping between objects' visible pyramids, 3D visible volume is formulated as:

$$VV_S = V_S - \sum_{i=1}^{N_{obj}} \sum_{j=1}^{N_{surf}} (V(PVP_i^j) - V(VP_i^j) + IHV_i^j) \qquad (5)$$

The same analysis holds true for multiple overlapping objects, adding the IHV between each two consecutive objects.

In Figure 9, we demonstrate the case of three buildings with overlapping. The invisible surfaces are bounded with dotted lines, while the projected visible surfaces to the overlapped building are colored in gray. In order to calculate the visible volumes from a viewpoint, *IHV* between each two buildings must be added as a visible volume, since it is already omitted at the previous step as an invisible volume.

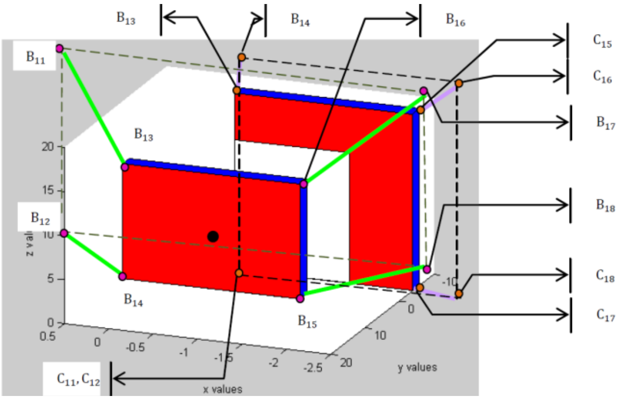


Figure 8. Invisible Volume $V(PVP_i^j) - V(VP_i^j)$ colored in purple and green arrows for each building. PVP of the object close to viewpoint colored in black, colored with pink circles and the far object PVP colored with orange circle

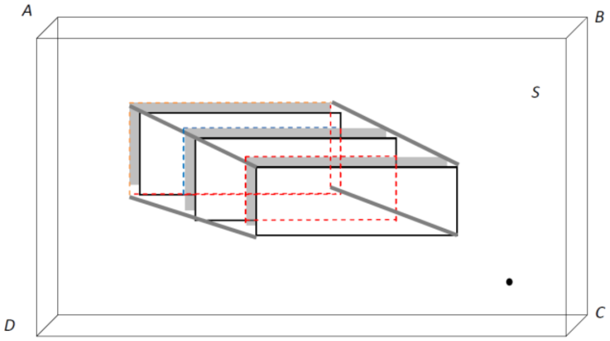


Figure 9. Three overlapping buildings. Invisible surfaces bounded with dotted lines, projected visible surfaces of the overlap building colored in gray

C. Simulations

In this section, we demonstrate the SVC method of estimating the number of clusters based on pedestrians' mobility datasets. For each pedestrian's location datasets, we analyze the 3D visible volumes inside bounding volume *S*, defined as a 3D box.

Our datasets are based on the city of Melbourne's 24-hour pedestrian monitoring system (24PM). This system measures pedestrian activity at several Points of Interests (POI) with counting sensors. Pedestrian mobility datasets are available online with interactive maps, as seen in Figure 10(a), and can be downloaded for a specific date.

Our datasets include the number of pedestrians in each hour during the 2nd of July 2013, at different seventeen points of interest in Melbourne where counting sensors are located and defined as observation points.

Based on these datasets, we approximated the pedestrians' location using the well-known and common kinematic model for pedestrians presented by Hoogendoorn [12] etc. Based on this model, pedestrian 2D location can be estimated as:

$$x(t + \Delta t) = x(t) + V(t)\Delta t + w \qquad (6)$$

where *w* is a white noise of a standard Wiener Process which reflects the uncertainty in the expected traffic condition, described as Gaussian distribution.

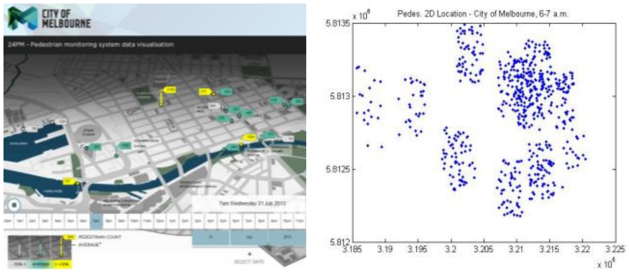
Pedestrian speed *V* can be divided into three major groups:

1. Fast: 1.8 meters per second
2. Standard: 1.3 meters per second
3. Slow: 0.8 meters per second

$$\begin{gathered} \mathrm{V(t)} \sim \mathrm{N}(\mu = 1.3, \sigma^2 = 0.5) \\ \mathrm{w} \sim \sqrt{\Delta \mathrm{t}}\mathrm{N}(0,1) \end{gathered} \qquad (7)$$

The kinematic model of a pedestrian is only a part of the estimation and prediction of his movement in an urban environment. For simplicity, we use only a kinematic model for a pedestrian's future location, since decision-making in this field is very complicated.

At time step t, pedestrian location $x(t)$, is taken from a specific POI from our dataset, and the estimated pedestrian location $x(t + \Delta t)$ can be computed. In our simulations we set Δt for five minutes. For example, pedestrians' 2D location in UTM coordination, using the Hoogendoorn model [12], etc., between 6-7 a.m., can be seen in Figure 10(b).


CITY OF MELBOURNE
24PM - Pedestrian monitoring system data visualisation
Pedes. 2D Location - City of Melbourne, 6-7 a.m.
x 10^6
5.8135
5.813
5.8125
5.812
3.185
3.19
3.195
3.2
3.205
3.21
3.215
3.22
3.225
x 10^5

(a) (b)

Figure 10. (a) City of Melbourne's 24-hour pedestrian monitoring system (24PM) – Online Visualization Map. (b) Pedestrians' 2D estimated location using the Hoogendoorn model [12] etc. between 6-7 a.m.

Each of pedestrian locations is processed as a viewpoint for estimating the number of clusters from spatial visibility aspects. The 3D visible volumes computation presented in the previous section are applied for computing $\mathrm{T_k}$, as described in Section II-A.

At each POI, we set the reference dataset of the pedestrian location distributed uniformly around the POI location, where the reference dataset size is the same one as the original dataset for the same POI, computing T_k^* .

We set the possible number of clusters from one to ten, demonstrating the SVC method. The number of clusters based on visible volumes analysis per day hour is presented in Figure 11.

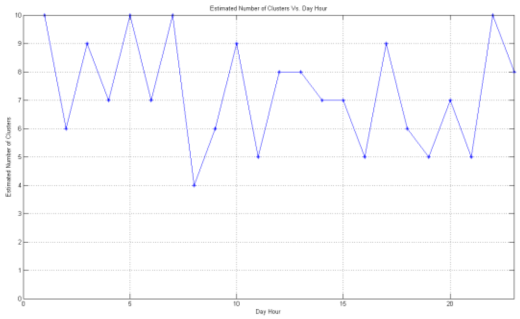


Figure 11. Number of Clusters for each Hour of 2/7/2013 Using SVC

As we can see in Figure 11, there is a correlation between the number of clusters and the pedestrians' mobility behavior. The number of clusters is close to the maximum (ten clusters in our case) during 6-9 AM, as can be predicted due to pedestrians' mobility while going to work. The number of clusters drops to a figure between eight to four clusters during the midday hours, and climbs again during nigh hours. More incentives analyzing pedestrians' mobility patters are presented in the next section.

III. Analyzing Pedestrians' Mobility Datasets

A. *Control Points*

In this section, we analyze pedestrians' mobility datasets during one day, estimating the number of clusters by using the SVC outcome, which is based on visibility analysis. Upon that, we use the K-means clustering method.

K-means clustering intends to partition n objects into k clusters, where each object belongs to the cluster with the nearest mean. The centroid of all objects in each cluster is set as control point. This method produces exactly k different clusters, where k is predefined from the SVC method. The objective of K-means clustering is to minimize total intra-cluster variance, or the squared error function.

K-means algorithm is a heuristic algorithm which depends on initial cluster. K-means can be very slow to converge, but in practice can be handling as polynomial convergence case which is the same case for our data sets [19].

By using K-means and SVC method, control points location can be seen in Figure 12.

It can be noticed in Figure 12 that, in some cases, the geometric location of the sensor location is separated into two different clusters. Our maximal number of clusters is set to ten, whereas there are seventeen sensors. We set the maximal number of clusters to be smaller than the number of sensors on the scene. One of the major contributions of our work, related to the adaptive clustering capability, is separating datasets into a different clustering and setting the control points from a visibility aspect. Moreover, control point location should cover more than one area, as can be seen in Figure 12, and also depends on pedestrians' mobility during this hour, as can be seen in the next sub-section.

Video simulations showing control points locations using K-means clustering and SVC methods are available in [16].

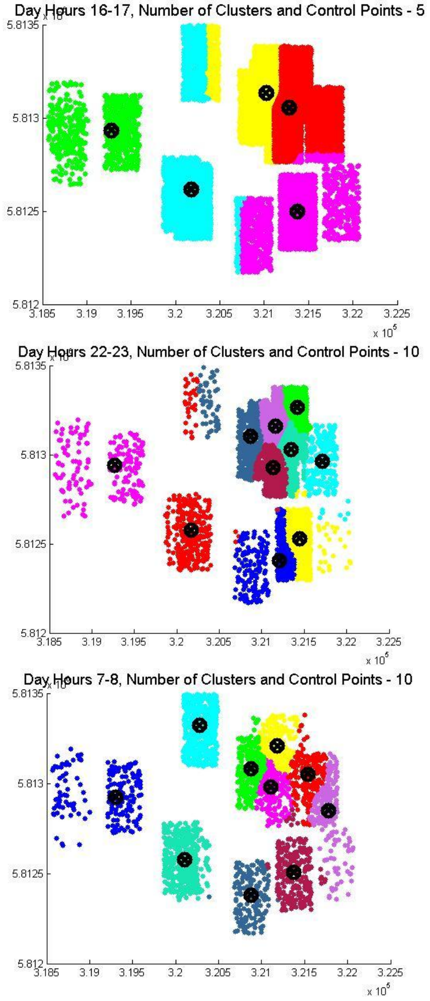


Figure 12. Control Points Location and Clusters Presentation during Each Hour in a Day. Control points are marked with black circles. Pedestrians' mobility Clustered in different colors

B. *Time Zones*

In this section, we concentrate on learning pedestrians' patterns for setting optimal control points, i.e., control points for each time zone.

We divide the day into four time zones for efficient pedestrian monitoring: Morning hours (movement to work) – 6 - 9 AM; Mid-Day Hours (between morning and afternoon) – 10 AM to 16 PM; Afternoon hours (back from work and activity hours) – 17- 20 PM.; Night hours 20 - 23 PM.

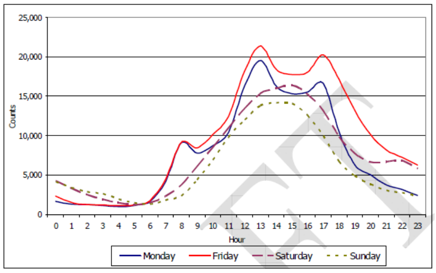


Figure 13. Pedestrian Activity Analysis [14]

The suggested division of time zones partition can also be seen clearly in an official pedestrian monitoring report of the city of Melbourne [14], in Figure 13. The number of pedestrians counted by the monitoring system rises at the suggested time zones. In order to get reliable and comprehensive results regarding pedestrian mobility patterns, we tested a full month's (July 2013) dataset, analyzing each day for twenty-four hours.

Based on the average estimated number of clusters using SVC on these datasets, we found out that the number of optimal control points during these time zones is: Morning hours - Nine control points; Mid-Day Hours - Six control points; Afternoon hours - Seven control points; Night hours - Eight control points.

The localization of the optimal control points and the number of clusters for each time zone can be seen in Figure 14.

It can be seen that in the different time zones, three optimal control points and their cluster division are almost identically marked with arrows and numbers in Figure 14. Four optimal control points with similar clustering can be seen in three time zones in Figure 14. These results can be applicable for personal-security and homeland security application in urban environments, localizing forces and sensors for optimal monitoring during a day's hours.

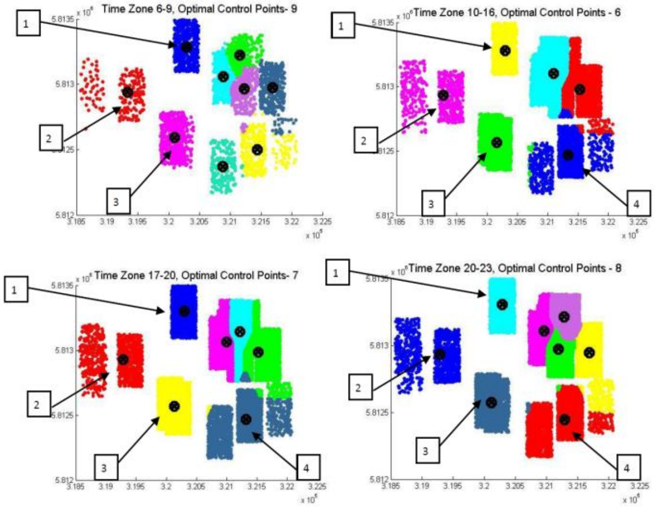


Figure 14. Optimal Control Points Location in Four Time Zones. Optimal Control points marked with black circles. Pedestrians' mobility Clustered in different colors

IV. CONCLUSIONS

In this paper, we presented a unified spatial analysis defining the number of clusters in a dataset based on an analytic visibility analysis, SVC.

The SVC method is based on fast and efficient 3D visible volumes computation. Estimating the number of clusters is based on minimum normalized visible volumes to reference datasets distributed uniformly inside bounding volume S. We demonstrated the SVC by using datasets from the city of Melbourne's 24-hour pedestrian monitoring system (24PM).

In the second part of this research, based on the SVC-estimated number of clusters, we analyzed pedestrian's mobility behavior, setting control points during a day's hours and dividing a day's hours into four time zones. We found a correlation of several optimal control points in different time zones.

Based on similar spatial analysis in other urban scenes, one can set optimal control points for various applications, such as entertainment events that can be efficiently visible at such points, or monitoring crowds' pedestrians from these control points in emergencies, equipped with medical assistance. Such results are also applicable for personal security and homeland security applications in urban environments, localizing police forces and sensors for optimal monitoring at different hours in a day.

V. REFERENCES

[1] P. Arabie and L. J. Hubert, "An Overview of Combinatorial Data Analysis", in Arabie, P., Hubert, L.J., and Soete, G.D. (Eds.) Clustering and Classification, 1996, pp. 5-63, World Scientific Publishing Co., NJ.

[2] A. Borgers and H. Timmermans, "A model of pedestrian route choice and demand for retail facilities within inner-city shopping areas", Geographical Analysis, Vol. 18, No. 2, 1996, pp. 115-128.

[3] R. B. Calinski and J. Harabasz, “A Dendrite Method for Cluster Analysis”, Communications in Statistics ,vol. 3, 1974, pp. 1–27.

[4] Y. Doytsher and B. Shmutter, "Digital Elevation Model of Dead Ground", Symposium on Mapping and Geographic Information Systems (Commission IV of the International Society for Photogrammetry and Remote Sensing), Athens, Georgia, USA, 1994.

[5] F. Durand, "3D Visibility: Analytical Study and Applications", PhD thesis, Universite Joseph Fourier, Grenoble, France, 1999.

[6] V. Estivill-Castro and I. Lee, "AMOEBA: Hierarchical Clustering Based on Spatial Proximity Using Delaunay Diagram." In Proceedings of the 9th International Symposium on Spatial Data Handling. Beijing, China, 2000.

[7] O. Gal and Y. Doytsher, "Fast and Accurate Visibility Computation in a 3D Urban Environment", in Proc. of the Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services, Valencia, Spain, 2012, pp. 105-110, [accessed February 2014].

[8] O. Gal and Y. Doytsher, "Fast Visibility Analysis in 3D Procedural Modeling Environments", in Proc. of the, 3rd International Conference on Computing for Geospatial Research and Applications, Washington DC, USA, 2012.

[9] O. Gal and Y. Doytsher, "Fast Visibility in 3D Mass Modeling Environments and Approximated Visibility Analysis Concept Using Point Clouds Data", Int. Journal of Advanced Computer Science, IJASci, Vol 3, No 4, April 2013, ISSN 2251-6379, [accessed February 2014].

[10] O. Gal and Y. Doytsher, "Fast and Efficient Visible Trajectories Planning for Dubins UAV model in 3D Built-up Environments", Robotica, FirstView, Article pp 1-21 Cambridge University Press 2013 DOI: http://dx.doi.org/10.1017/S0263574713000787, [accessed February 2014].

[11] A. Gordon, Classification (2nd ed.), London: Chapman and Hall/CRC Press, 1999.

[12] S. P. Hoogendoorn and P. H. L. Bovy, "Microscopic pedestrian way finding and dynamics modelling," In Schreckenberg, M., Sharma, S.D. (eds.) Pedestrian and Evacuation Dynamics. Springer Verlag: Berlin, 2001, pp. 123-154.

[13] Melbourne City: http://www.pedestrian.melbourne.vic.gov.au/#date=31-08-2013&time=9 [accessed February 2014].

[14] Melbourne Report: https://docs.google.com/file/d/0B380gpj_lbU-dnRaRTFXdlh5Znc/edit [accessed February 2014].

[15] H. Plantinga and R. Dyer, "Visibility, Occlusion, and Aspect Graph," The International Journal of Computer Vision, vol. 5, 1990, pp. 137-160.

[16] https://sites.google.com/site/orenusv/home/svc [accessed February 2014].

[17] T. Schelhorn, D. Sullivan, M. Haklay, "STREETS: An agent-based pedestrian model," http://www.casa.ucl.ac.uk/streets.pdf. , 1999, [accessed February 2014].

[18] R. Tibshirani, G. Walther, T. Hastie, “Estimating the Number of Clusters in a Dataset via the Gap Statistic,” Journal of the Royal Statistical Society, Ser. B, vol. 32, 2001, pp. 411–423.

[19] A. Vattani, "K-means Requires Exponentially Many Iterations Even in the Plane", Discrete and Computational Geometry, vol. 45 (4), 2011, pp. 596–616. DOI:10.1007/s00454-011-9340-1, [accessed March 2014].

A Semantic Analysis of Moving Objects Using as a Case Study Maritime Voyages from Eighteenth and Nineteenth Centuries

Helbert Arenas, Benjamin Harbelot, Christophe Cruz
Department of Informatics
University of Burgundy
Dijon, France 21078
helbert.arenas@checksem.fr
benjamin.harbelot@checksem.fr
christophe.cruz@u-bourgogne.fr

Abstract—**In this paper, we present a spatial model designed to extract knowledge from the tracks of moving objects. The model uses a *perdurantism* approach, implemented using Semantic Web technologies. In order to show the capabilities of the model, we employ it to handle a large dataset composed of historic maritime records. By using Semantic Web tools, we are able to implement rules and identify user defined patterns. Although there are limitations due to currently available tools, our results are promising.**

Keywords–spatial modeling; moving objects; semantics.

I. INTRODUCTION

The two main philosophical theories for the representation of objects evolving along time are: *endurantism* and *perdurantism*. The first one, *endurantism*, considers objects as three dimensional entities that exist wholly at any given point of their life. On the other hand, *perdurantism*, also known as the four dimensional view, considers that entities have temporal parts, *timeslices* [1]. Each *timeslice* is a partial representation of the object, valid only for a specific period or point of time. A complete representation of the object along time is the result of aggregating all its *timeslices*. From a designer point of view, the *perdurantism* approach offers advantages over the *endurantism* one, by allowing richer representations of real world phenomena [2].

Most of current Geographic Information Systems (GIS) tools use a *snapshot* approach to study spatial systems. With this type of tools, it is difficult to analyse dynamic phenomena with spatial-temporal dimensions. An alternative to traditional data management approaches is the Semantic Web. This is a set of standards that enable sharing data and semantics of the data on the web. Using Semantic Web related technologies, it is possible to develop data models called ontologies specifically designed for reasoning and inference with software mechanisms. Ontologies allow for any given domain, the representation of relevant high level concepts as well as their properties and the relationships between concepts and entities. In this research, we use Semantic Web technologies to develop the “continuum model”, an ontology that allows us to represent diverse dynamic entities and analyse their relationships along time. Traditionally, ontologies are static in the sense that the information represented in them does not change in time or space. However, previous research such as [1] [3] [4] and [5] aim to fill this gap, by developing ontologies that use a *perdurantism* approach to handle dynamic entities. The focus of [4] and [5] is on evolving entities represented in space as areas. In this paper, we propose an extension of the previous research, changing the focus to moving entities. In this paper, we use historical maritime records as a test bed. In Section II, we identify other works in this field. Section III provides a description of the dataset we use. We describe our model and how we implement it in Section IV. Finally, in Section V, we present our conclusions and indicate our future research in this field.

II. RELATED RESEARCH

Currently, new datasets containing large amounts of tracking data are becoming available. These datasets contain the recorded position of a travel entity while it moves in space. In order to extract knowledge from this information, a new set of tools and algorithms are being developed in the research world. In Parent et al., 2013, the authors present a survey on current approaches and techniques for the definition of *semantic trajectories* within the field of *data mining*. This paper describes techniques to create trajectories from raw data, to add semantics to the trajectories and to extract knowledge [6].

A raw *trajectory* comprises the record of the position of a moving entity, during a time interval, in which this entity moves for some meaningful purpose. In some domains, the identification of the begin and end of the trajectory is evident, while in others this might require some domain specific criteria. The positions and intervals allow us to identify possible *stops* along the trajectory. Additional information can be obtained by linking the trajectory to external datasources. For instance, comparing the track record of a person with a set of places of interest of a city, could tell us what places has the person visited [6].

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

By analyzing the trajectory, it is possible to identify stops and elements on the route. Then, we can add *annotations* creating in this way a *semantic trajectory*. An analysis of the trajectory can also lead to the identification of a particular behaviour of the moving entity. For instance, by analyzing the elements on the route of a tracked person, we could distinguish a tourist from a pizza delivery service [6].

An implementation of the previously described concepts is presented in Spaccapietra et al., 2008 [7]. Here, the authors implement their approach with a relational Database Management System (DBMS), using as a case study data from the annual migrations of white storks *(Ciconia ciconia)*.

An approach that puts more focus on Description Logics (DL) based tools is presented in Yan et al., 2008. Here, the authors propose the use of three different ontologies to address the study of moving entities. The first one is caled *Geometric Trajectory Ontology*. This ontology defines the basic concepts for the spatio-temporal definition of the trajectory. Using the elements defined in this ontology, we can specify temporal points, areas, lines etc. The second one is called *Geography ontology*. In this ontology, the authors describe natural and artificial features of interest for the specific domain. Finally, the third one is called *Application Domain Ontology*. In this ontology, the authors define higher level concepts for specific domains. In Yan et al., 2008, the authors test their ideas using data from cars equipped with GPS devices. The data is loaded into a commercial relational DBMS with support for ontological data [8].

In Yan et al., 2011, the authors introduce the Semantic Middleware for Trajectories (SeMiTri). This software is designed to create *annotations* by analyzing the geometric properties of the trajectory and linking it to background geographic and application specific data. The proposed system has three parts: 1) *Trajectory computation layer*, here the raw GPS data is cleaned, raw trajectories are identified and each trajectory is divided into trajectory episodes. 2) *Semantic annotation layer*, to link the trajectory to areas it crosses, road networks. It also estimates probabilities of association between *stops* and geographic features using a Markov model algorithm. 3) *Semantic trajectory analytics layer*, here are the components of the system that compute statistics, and store obtained information. An additional component is the Web Interface, designed to allow the user to define queries and visualize results [9].

An example of a *perdurantism* approach for spatial dynamics is presented in Harbelot, Arenas & Cruz, 2013b. In this research, the authors introduce the *continuum* model, a methodology that can successfully represent the changes of entities in space and property values along time [5]. However, this approach focuses on entities represented as areas, making the approach not well suited for moving entities represented as points. In this paper, we further define the *continuum* model to represent traveling entities. In the next sections, we will describe the datasets and the model we propose.

III. Datasets

In this paper, we present a methodology to model spatial moving objects using a Semantic Web approach. Our moving objects are ships from the eighteenth and nineteenth centuries, whose positions have been recorded in logbooks. Between 2001 and 2003 the European Union funded the project: Climatological Database for the World's Oceans (CLIWOC). A product of this project is the digitized version of logbooks of pre 1854 voyages of English, Spanish, Dutch and French ships [10][11].

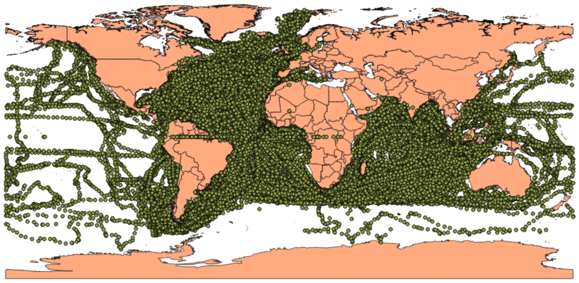


Figure 1. CLIWOC dataset, points represent logbook records.

The CLIWOC datasets became available online in 2003 [12]. The dataset contains 280280 records. Each record on the logbooks contains the position of the ship, as well as meteorological observations (temperature, wind speed, atmospheric pressure, etc.). Figure 1 depicts a map with the recorded logbook observations. Due to technological limitations, some of the positions recorded might have spatial errors, as can be noted by observing the positions recorded within land masses.

The creators of CLIWOC processed more than 3000 logbooks that represent more than 5000 voyages through world's oceans. Entries in the logbooks were done at noon. For each entry, the officer in charge registered the observed climatic conditions at the time. The CLIWOC dataset allows scientists to study weather patterns in the eighteenth and nineteenth centuries. By analyzing the records, it is possible to infer the evolution of phenomena such as the Nino Southern Oscillation or the North Atlantic Oscillation. Surprisingly, the dataset represent less than 10% of data available from original documents [12].

The climatic measurements in the logbooks were taken using archaic methodologies. For instance, in the case of the wind force, the measurements were recorded using old terms no longer in use. It is therefore necessary to translate them to modern units for study purposes. This problem was solved by the CLIWOC team with a dictionary that allows the translation from each archaic language vocabulary into Beaufort scale terms [12].

The CLIWOC dataset has been previously used to study climatic patterns. However, as noted by [13], the dataset can be used to other fields. For instance, it is possible to study the spread of technology. Maritime chronometers were popular on ships of the East Indian Company before they became common on Royal Navy vessels. With better location techniques, sailors were able to modify their routes and take advantage of more favourable wind patterns, which meant a modification on the routes. Using the logbooks, it is also possible to study the evolution of the crew health at ships of different countries, along time. From the end of XVIII century, improvements in

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

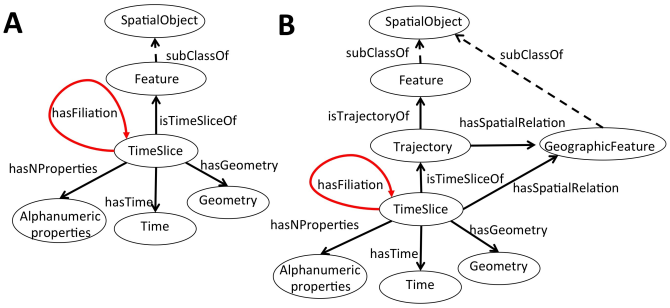


Figure 2. A)The *Continuum* model as introduced in [5]. B) Modified version of the *Continuum* model adapted for moving objects.

ventilation, diet and hygiene reduced the amount of diseases in maritime travels. Health information was recorded in logbooks of all the countries in the study, allowing scientist to make comparable studies. Another possible research topic is the influence of seasonal weather patterns on travels. For instance, British and Dutch ships were strongly affected by seasonal weather patterns when travelling to the Indian Ocean. Because of this, it was very unusual to see ships of these nationalities sailing in the South Atlantic around November and December [13].

In order to enhance the knowledge we can extract from the CLIWOC dataset, we created a *Geography Ontology* following the approach suggested by [8]. The geographic entities in this ontology are the seas and oceans of the world. We obtained this information from a high resolution dataset introduced in [14]. The creators of this dataset based their work on the document S-23, titled *Limits of Oceans and Seas* published by the International Hydrographic Organization (IHO) in 1953 [15]. The vector dataset, although based on an official document, it is not an official document itself. The vector dataset has a very high resolution coastline that is not necessary in our research. In order to facilitate our spatial operations, we simplified it using GeoTools.

IV. Proposed Model

In Harbelot et al., 2013a and Harbelot et al., 2013b [4] [5], the authors introduced the *continuum* model, an approach well suited to represent dynamic entities represented spatially as areas. Figure 2A depicts the continuum model as described in [5]. This model follows a *perdurantism* approach. It creates multiple ephemeral representations (*timeslices*) to depict a dynamic entity. Each *timeslice* is valid for a determined time interval. A *timeslice* has four components: 1) An identity, linking it to the object it represent. 2) A set of properties with alphanumeric values, representing different characteristics of the object. 3) A time component that indicates the valid period of time for the *timeslice* and 4) A geometric component, the ephemeral spatial representation of the entity. The model creates a new *timeslice* everytime there is a change in the geomety, the identity or in the alphanumeric properties. It is then possible to establish a filiation relationships between a newly created *timeslice* and the one that originated it.

The *continuum* model needs to be modified in order to deal with tracking information of travelling entities. Figure 2B depicts our upgraded version of the *continuum* model. In the new model, we have the moving objects as instances of the class *Feature*. We split the movements of the objects into *Trajectories*, which are semantic units with a defined *start* and *end* spatio-temporal points. The *Trajectory* itself is composed of a set of *timeslices*, with the same components as in the original version of the *continuum* model.

Following the approach suggested by [8] we developed a *Geographic Ontology* composed of *GeographicFeatures*. This ontology allows us to extract new knowledge from the tracking information.

A. Model Specification

In our model, we define the following concepts:

- *Temporal Points* We can think of the temporal domain as a linear structure $\mathcal{T}$ composed of a set of temporal points $\mathcal{P}$. The components of $\mathcal{P}$ follow a strict order $<$, which forces all points between two temporal points t_1 and t_2 to be ordered [16].

$$\mathcal{P} \sqsubseteq \top \tag{1}$$

- *Geometries* A set of coordinates that define points, lines, curves, surfaces and polygons.

$$\mathcal{G} \sqsubseteq \top \tag{2}$$

- *Spatial Objects* This class represents any entity with geometric representation.

$$\mathcal{O} \equiv \exists hasGeometry.\mathcal{G} \tag{3}$$

- *Spatio-Temporal Point* We use this class to define points with a spatial and temporal representations.

$$\mathcal{TP} \equiv \exists hasGeometry.\mathcal{G} \sqcap \exists hasTime.\mathcal{P} \tag{4}$$

- *Moving Features* This class is used to represent entities that change their positions along time.

$$\mathcal{MF} \sqsubseteq \mathcal{O} \tag{5}$$

The movement of the feature is divided into semantical units called trajectories.

$$\mathcal{MF} \equiv \exists hasTrajectory.\mathcal{TR} \tag{6}$$

- *Trajectories* This class is used to represent semantical units of movement, with a defined start and end spatio-temporal points.

$$\begin{aligned}\mathcal{TR} \equiv &\exists hasTimeSlice.\mathcal{TS} \sqcap \exists isTrajectoryOf.\mathcal{MF} \\ &\sqcap \exists hasStart.\mathcal{TP} \sqcap \exists hasEnd.\mathcal{TP}\end{aligned} \tag{7}$$

- *TimeSlice* This class is used to depict a partial representation of a moving entity. Each *timeslice* has a defined geometric and temporal representations. It also has an identity component that links it to a specific *trajectory* and a set of alphanumeric properties $(\mathcal{TS})$ that represent diverse characteristics of the entity at the specific point of time defined by its temporal dimension.

$$\begin{aligned}\mathcal{TS} \equiv & \exists hasGeometry.\mathcal{G} \sqcap \exists hasTime.\mathcal{P} \\ & \sqcap \exists \overline{\mathcal{TS}} \sqcap \exists isTimeSliceOf.\mathcal{TR}\end{aligned} \tag{8}$$

- *Geographic Entity* Following the approach suggested by [8], we implement a external ontology, composed of geographical entities GE. By combining the track of moving objects with these external geographic entities, we can improve the knowledge extraction. In our test study, we use a seas and oceans dataset based on a document published by International Hydrographic Organization (IHO) in 1953 [14].

$$\mathcal{GE} \sqsubseteq \mathcal{O} \tag{9}$$

Then, we can create individuals for each of the classes/concepts. For instance, $\mathcal{TS}(ts)$ indicates that ts is an individual of type *timeslice* ((TS)). Following the definition of the class TS, we know that the individual ts has a geometric component ts_g which is an instance of the class $\mathcal{G}$, then $\mathcal{G}(ts_g)$. The temporal dimension of ts is represented by ts_t, which is an instance of the class Temporal Points ($\mathcal{P}(ts_t)$). The individual ts is linked to a trajectory (ts_{tr}), and through the *trajectory* to a specific individual of the type *Feature* , and in this way it has a defined identity. We represent the alphanumeric properties that describe the characteristics of the timeslice $ts1$ as $ts1_{\overline{ts}}$.

Using the temporal and identity components of the *timeslices*, it is possible to identify a sequence, and in this way establish a *filiation* relationship between two *timeslices*. For the relationship to exist, both *timeslices* must belong to the same trajectory, and there should no exist other *timeslice* of the same trajectory occurring in between.

$$\begin{aligned}& after(ts1_t, ts2_t) \\ \wedge\neg\exists(ts \in \mathcal{TS})|(ts1_t < ts_t < ts2_t) \wedge & (ts1_{tr} = ts2_{tr} = ts_{tr}) \\ & \rightarrow hasFiliation(ts1, ts2)\end{aligned} \tag{10}$$

Then, we can compare *timeslices* that hold *filiation* relationships. For instance, we can calculate the speed of the moving entity for the interval defined between them as the following:

$$\begin{aligned}speed(ts1, ts2) \equiv & \left(\frac{distance(ts1_g, ts2_g)}{timeDiff(ts1_t, ts2_t)}\right) \\ & |hasFiliation(ts1, ts2)\end{aligned} \tag{11}$$

Using the *filiation* relationships and the *speed* calculation, we can identify episodes with unusual behaviours. For instance, a very low speed between two timeslices might suggest a *stop*. On the other hand, an unusually high speed might suggest errors in the geometric component of one of the timeslices, requiring further attention by the researcher. To identify unusually high speeds, we can calculate the statistics of the speeds of a *trajectory*.

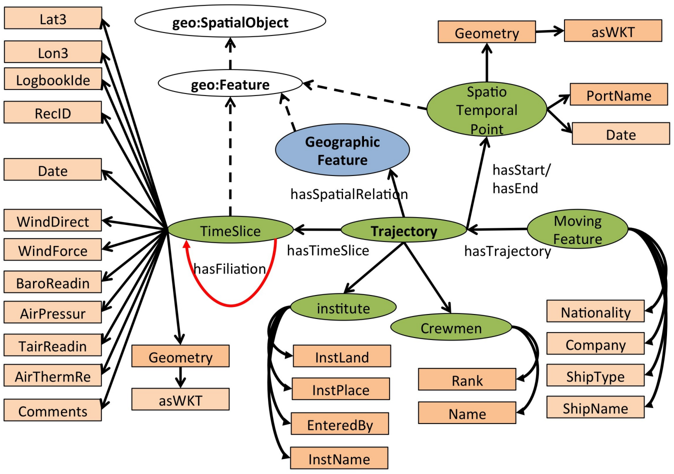


Figure 3. Classes and properties defined to model the CLIWOC dataset using the *Continuum* model

$$\begin{aligned}(speed(ts1, ts2) - mean(tr_{speed})) > \lambda(\sigma(tr_{speed})) \\ \wedge(hasFiliation(ts1, ts2)) \\ \rightarrow suspiciousFiliation(ts1, ts2)\end{aligned} \tag{12}$$

Where tr is a given trajectory, $mean(tr_{speed})$ depicts the average speed of the trajectory tr. The standard deviation of speed values is represented by $\sigma(tr_{speed})$, while λ represents a scalar value, by default 3.

We can later identify specific timeslices with geometries that require further attention by the researcher.

$$\begin{aligned}\exists suspiciousFiliation(ts1, ts) \\ \wedge\exists suspiciousFiliation(ts, ts2) \\ \rightarrow suspiciousTimeSlice(ts)\end{aligned} \tag{13}$$

A suspicious timeslice only indicates that the speed required to reach this point is unusually high. However, it does not prove that the position of the timeslice is incorrect.

We can extract knowledge by establishing spatial relationships between trajectories with an external geographic entities

$$\begin{aligned}\exists ts|hasTimeSlice(tr, ts) \wedge (Intersect(geo_g, ts_g)) \\ \vee(Within(geo_g, ts_g)) \vee (DistanceWithin(geo_g, ts_g, d))) \\ \rightarrow hasSpatialRelation(tr, geo)\end{aligned} \tag{14}$$

where ts is a timeslice ($TS(ts)$), tr is a trajectory ($\mathcal{TR}(tr)$) and geo is an geographic entity ($\mathcal{GE}(geo)$).

Figure 3 depicts a detailed representation of the continuum model with the CLIWOC dataset. The classes are represented as ellipses, while the values for the various properties are represented as squares.

ISBN: 978-1-61208-326-1

TABLE I. SUMMARY OF THE ENTITIES UPLOADED INTO THE TRIPLESTORE

Count	Ontology Class	Description
280280	$\mathcal{TS}$	Records from logbooks, each record is represented as a timeslice and has a 0 dimensional spatial representation (point). Each timeslice has a link to a trajectory. Each timeslice has also climatic observations and a date entry property.
5198	$\mathcal{TR}$	Voyages, as identified in the original CLIWOC dataset. Each voyage has a departure and end spatio-temporal point. Each voyage is linked to the ship it corresponds.
1261	$\mathcal{MF}$	Ships, are the moving features in the ontology. Each ship has the properties : hasShipName, hasNationality, hasCompany and hasShipType. In some cases, the value property was not available on the dataset.
105	$\mathcal{GE}$	Oceans/Seas, represented as polygons. The dataset covers the whole world. It required a simplification process before being uploaded. We simplified the coastlines and removed the islands.

B. Implementation of the model using the CLIWOC dataset

The knowledge in the Semantic Web is represented using standards such as the Resource Description Framework (RDF) [17] and Web Ontology Language (OWL) [18], while the traditional ontology query language is the Protocol and RDF Query Language (SPARQL) [19]. It is possible to query spatial information stored in an ontology, using GeoSPARQL [20]. This is a set of standards from the Open Geospatial Consortium (OGC). In our research, we decided to implement our model and store it in a Parliament triplestore. We opted for Parliament due to its support for GeoSPARQL. In order to upload the datasets to the triplestore, we developed a program in Java using the Jena and GeoTools libraries. Table I contains a summary of the entities uploaded to the triplestore.

Due to performance reasons, we opted to perform certain operations outside the triplestore, using a Java application instead. For instance, the CLIWOC dataset uses geographic coordinates, therefore distance calculation is not a trivial task. We decided to perform this operation with Java, because we could have better control over the results. It was also possible to test different formulas and make performance comparisons.

Once the data was uploaded into the triplestore, it was possible to identify the *filiation* relationship as defined in equation 10. The following query is a SPARQL translation of equation 10. It detects the filiation relationships between timeslices corresponding to the trajectory $abc : trajectory\_5198$.

```
INSERT
{?tsParent abc:hasFiliation ?tsChild.}
WHERE {
abc:trajectory_5198 abc:hasTimeSlice ?tsParent.
abc:trajectory_5198 abc:hasTimeSlice ?tsChild.
?tsParent abc:hasDate ?ParentDate.
?tsChild abc:hasDate ?ChildDate.
NOT EXISTS{
abc:trajectory_5198 abc:hasTimeSlice ?tsX.
?tsX abc:hasDate ?XDate.
FILTER
((?ParentDate <?XDate) &&
(?XDate <?ChildDate))}
FILTER
((?tsParent!=?tsChild) &&
(?ParentDate<?ChildDate))}
```

According to [13], *British* and *Dutch* vessels on route to the Indian Ocean were affected by seasonal wind patterns. Because of this, vessels of these nationalities were not seen in the Southern Atlantic in the months of November and December. Using the model we can identify these unusual voyages of *British* ships as:

$$\begin{aligned}\mathcal{GE}(geo), \mathcal{TR}(tr), \mathcal{TS}(ts)\\ |(geo_{name} = `SouthAtlantic')\wedge\\ (\exists ts|hasTimeSlice(tr, ts))\\ \wedge(Within(ts_g, geo_g))\\ \wedge(hasMonth(ts_t, `November')\vee\\ hasMonth(ts_t, `December'))\\ \rightarrow UnusualTrajectory(tr)\end{aligned} \tag{15}$$

Equation 15 can be translated in SPARQL as:

```
INSERT
{?t a abc:UnusualTrajectory}}
WHERE {
?t a abc:Trajectory.
?m a abc:MovingFeature.
?m abc:hasShipNationality "British".
?m abc:hasTrajectory ?t.
?t abc:hasTimeSlice ?ts.
?ts abc:hasDate ?tDate.
?tDate xsd:Month(?tDate)
?ts geo:hasGeometry ?tsGeo.
?tsGeo geo:AsWKT ?tsWKT.
?g a abc:GeographicFeature.
?g abc:hasGeographicName "South Atlantic".
?g geo:hasGeometry ?gGeo.
?gGeo geo:AsWKT ?gWKT.
FILTER((geof:sfIntersects(?gWKT,?tsWKT))&&
((str(month(?tsDate))="11")||
(str(month(?tsDate))="12")))}
```

The result of this query allows us to identify trajectories of ships of *British* nationality that in the opinion of an expert on the field, follow an unusual trajectory pattern [13].

By using an ontology, we are able to easily identify specific patterns on the data, enabling scientist to better understand large datasets containing records of moving entities.

V. CONCLUSION

In this paper, we present an approach to analyse historical maritime records. Our approach allows the knowledge discovery in large datasets, enabling scientist to identify patterns that might be hidden due to the large size of the dataset.

Our starting point is very basic raw data, from which we produce more sophisticated constructions that allow a better understanding of the events depicted in the dataset.

Currently, we use a state of the art triplestore as our data repository, which gives us several advantages: 1) We can maintain formal defined relationships between objects. 2) It allows us flexibility, we can easily define new properties and relations between entities and concepts, and 3) It allows us to operate in large datasets in triple format.

At the moment, we use Parliament, a triplestore that supports GeoSPARQL, allowing us to perform spatial analysis without additional software. GeoSPARQL is an OGC standard that comprises a set of functions that extend SPARQL. Thanks to using OGC standards, our approach can be deployed in alternative OGC compliant environments.

We plan to continue our research in the field of semantic modelling of dynamic entities. An interesting field of research is the definition of semantic rules. At the moment, there are two submissions to the World Wide Web Consortium (W3C) [21] that aim to work in this field: SPARQL Inference Notation (SPIN) [22] and Semantic Web Rule Language (SWRL) [23]. However, none of them is yet an official W3C standard. In the future, we plan to explore the rule definition option, sticking to accepted standards, securing in this way the extensibility of our work.

Acknowledgment

This research is supported by: 1) Conseil régional de Bourgogne. 2) Direction Générale de l'Armement, see: http://www.defense.gouv.fr/dga/.

References

[1] C. Welty and R. Fikes, "A reusable ontology for fluents in owl," in Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006). Amsterdam, The Netherlands, The Netherlands: IOS Press, 2006, pp. 226–236. [Online]. Available: http://dl.acm.org/citation.cfm?id=1566079.1566106

[2] M. M. Al-Debei, M. M. al Asswad, S. de Cesare, and M. Lycett, "Conceptual modelling and the quality of ontologies: Endurantism vs. perdurantism," International Journal of Database Management Systems IJDMS, vol. 4, no. 3, 2012.

[3] M. OConnor and A. Das, "A method for representing and querying temporal information in owl," in Biomedical Engineering Systems and Technologies, ser. Communications in Computer and Information Science, A. Fred, J. Filipe, and H. Gamboa, Eds. Springer Berlin Heidelberg, 2011, vol. 127, pp. 97–110. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-18472-7_8

[4] B. Harbelot, H. Arenas, and C. Cruz, "A semantic model to query spatialtemporal data," in Information Fusion and Geographic Information Systems (IF AND GIS 2013), ser. Lecture Notes in Geoinformation and Cartography, V. Popovich, C. Claramunt, M. Schrenk, and K. Korolenko, Eds. Springer Berlin Heidelberg, 2014, pp. 75–89. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31833-7_5

[5] ——, "Continuum: A Spatiotemporal Data Model to Represent and Qualify Filiation Relationships," November 2013, paper presented at the Fourth ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS) 2013, Orlando, Florida, USA.

[6] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan, "Semantic trajectories modeling and analysis," ACM Computing Surveys, vol. 45, no. 4, Aug. 2013, pp. 42:1–42:32. [Online]. Available: http://doi.acm.org/10.1145/2501654.2501656

[7] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot, "A conceptual view on trajectories," Data Knowledge Engineering, vol. 65, no. 1, Apr. 2008, pp. 126–146. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2007.10.008

[8] Z. Yan, J. Macedo, C. Parent, and S. Spaccapietra, "Trajectory ontologies and queries," Transactions in GIS, vol. 12, 2008, pp. 75–91. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9671.2008.01137.x

[9] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer, "SeMiTri: A framework for semantic annotation of heterogeneous trajectories," in Proceedings of the 14th International Conference on Extending Database Technology, ser. EDBT/ICDT '11. New York, NY, USA: ACM, 2011, pp. 259–270. [Online]. Available: http://doi.acm.org/10.1145/1951365.1951398

[10] R. Garcia Herrera, D. Wheeler, G. Konnen, F. Koek, P. Jones, and M. R. Prieto, "CLIWOC final report, UE contract EVK2-CT-2000-00090," 2004.

[11] R. Garcia-Herrera, G. Knnen, D. Wheeler, M. Prieto, P. Jones, and F. Koek, "CLIWOC: A Climatological Database for the World's Oceans," Climatic Change, vol. 73, no. 1-2, 2005, pp. 1–12. [Online]. Available: http://dx.doi.org/10.1007/s10584-005-6952-6

[12] R. Garcia-Herrera, G. P. Knnen, D. A. Wheeler, M. R. Prieto, P. D. Jones, and F. B. Koek, "Ship logbooks help analyze pre-instrumental climate," Eos, Transactions American Geophysical Union, vol. 87, no. 18, 2006, pp. 173–180. [Online]. Available: http://dx.doi.org/10.1029/2006EO180002

[13] C. Wilkinson, "The non-climatic research potential of ships' logbooks and journals," Climatic Change, vol. 73, no. 1-2, 2005, pp. 155–167. [Online]. Available: http://dx.doi.org/10.1007/s10584-005-6947-3

[14] D. Fourcy and O. Lorvelec, "A new digital map of limits of oceans and seas consistent with high-resolution global shorelines," Journal of Coastal Research, vol. 29, no. 2, 2013, pp. 471–477.

[15] I. H. Organization, "Limits of Oceans and Seas," 1953. [Online]. Available: hdl:10013/epic.37175

[16] A. Artale and E. Franconi, "A Temporal Description Logic for Reasoning About Actions and Plans," Journal of Artificial Intelligence Research, vol. 9, no. 1, Aug. 1998, pp. 463–506. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622797.1622809

[17] R. W. Group, "RDF primer," World Wide Web Consortium (W3C), 2014, accessed on March 2014. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

[18] O. W. Group, "OWL Web Ontology Language," World Wide Web Consortium (W3C), 2004, accessed on March 2014. [Online]. Available: http://www.w3.org/TR/2004/REC-owl-guide-20040210/

[19] S. W. Group, "SPARQL Query Language for RDF," World Wide Web Consortium (W3C), 2008, accessed on March 2014. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[20] OGC, "OGC GeoSPARQL - A Geographic Query Language for RDF Data," Open Geospatial Consortium (OGC), 2012, accessed on March 2014. [Online]. Available: http://www.opengeospatial.org/standards/geosparql#overview

[21] W3C, "World Wide Web Consortium institutional website," World Wide Web Consortium (W3C), 2014, accessed on March 2014. [Online]. Available: http://www.w3.org

[22] H. Knublauch, J. Hendler, and K. Idehen, "SPIN - overview and motivation," World Wide Web Consortium (W3C), 2011, accessed on March 2014. [Online]. Available: http://www.w3.org/Submission/spin-overview/

[23] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, "SWRL - A Semantic Web Rule Language Combining OWL and RuleML," World Wide Web Consortium (W3C), 2004, accessed on March 2014. [Online]. Available: http://www.w3.org/Submission/SWRL/

GIS Based Site Ranking using Neighbourhood Analysis and Comparison

Muhammad Irfan, Aleksandra Koj, Majid Sedighi and Hywel R. Thomas

Geoenvironmental Research Centre
School of Engineering, Cardiff University
Cardiff, UK
emails: {MuhammadI2, KojA, SedighiM, ThomasHR}@cf.ac.uk

Abstract—**This paper presents a Geographical Information Systems (GIS)-based toolkit, developed for site comparison and ranking that can be used to facilitate the decision making process in second stage of the site selection process. The toolkit has been developed as an analytical component of a multi-criteria spatial decision support system for geoenvironmental and geoenergy applications. The methodology adopted to develop this analytical module is based on a systematic comparison of the surrounding areas of each site in accordance with key environmental, socio-economic and public-health indicators. The sites are ranked based on the most favorable key indicators using a Criterion Sorting Mechanism (CSM) or Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). An application of the site selection toolkit is presented in relation to an unconventional geoenergy development. The application exercise deals with the ranking of a number of potential sites for coalbed methane recovery in Wales, UK. The locations of potential sites are first selected with respect to the gas resource (techno-economic viability). The toolkit is then used to select and rank the potential sites based on key environmental indicators, in the site's neighbourhood. The results of the site ranking using CSM and TOPSIS methods are compared and a number of scenarios are discussed. This approach of using a combination of site ranking methods along with the neighbourhood analysis reduces the risk of personal judgment and choice. The decisions on site selection can thus be evidenced on a quantified logic.**

Keywords-neighbourhood analysis; spatial decision support system; site selection; site ranking; coalbed methane.

I. Introduction

This paper describes the development of a Geographical Information Systems (GIS)-based toolkit for site ranking, based on an analysis of the surrounding areas. The toolkit can be used in the "second phase" of a site selection process, where the decision maker may need to choose from a number of potential sites or from a large suitable area.

The problem considered is one where after applying the "first phase" of a GIS-based site selection process, the decision maker is left with multiple choices either in the form of more than one equal potential sites (in vector format) or very large areas (in raster format).

Different GIS modelling techniques have been suggested in the literature for site selection process. These techniques are collectively known as Spatial Multi Criteria Decision Analysis (S-MCDA). For example, a combination of Analytical Hierarchy Process (AHP) and TOPSIS method has been suggested for municipal solid waste landfill site selection with a case study of Thrace region in Greece [1]. Weighted Linear Combination (WLC) in combination with fuzzy set theory has been applied for decisions on wind farm sites selection in Northwest Ohio [2]. AHP, fuzzy membership functions and Simple Additive Weighting (SAW) has been used together in a study to find the most suitable site for a tourist building a in the rural landscape of Hervás in Spain [3]. Similarly, a spatial decision tool for managed aquifer recharge has been presented in [4] that combine AHP, WLC and Ordered Weighted Averaging (OWA). A case study of this tool has been presented for the site selection of managed aquifer recharge in the Algarve Region of Portugal [4].

As described previously, at this stage, multiple sites may be obtained with equal potential. The suitable areas obtained in these studies are in the form of a raster map with relatively large areas, suitable for siting. For example, the area highlighted as a suitable area for aquifer recharge in [4], constitutes about 11.2% of the entire study area. Then, the suitable area was ranked on the basis of suitability score and the most highly suitable class reduced the area under consideration to 1% of the entire study area. At this stage the sitting decision is mainly based on the choice and expert knowledge of the decision maker(s).

To facilitate the decision making in the second phase, under the conditions described above, site ranking using neighbourhood analysis is presented in this paper. This approach can provide a quantified logic to select and rank the best site(s) from several candidate sites identified in the first level of the site selection process. Using the approach presented, inputs required from the decision maker and the risks associated with personal judgement and choice can be minimized.

An overview of neighbourhood analysis for site comparison and methods for site ranking is presented in Section II. In Section III, a description of the toolkit development and analytical components is presented. Tools and technologies used for the development of the toolkit are also highlighted. In Section IV, an application of the toolkit is presented which deals with the ranking of a number of hypothetical Coal Bed Methane (CBM) sites in Wales (UK). The toolkit has been used to rank the suitable sites in terms of key environmental indicators using both CSM and TOPSIS ranking methods. In Section V, the results of the application of the toolkit are discussed. Ranks generated from both the techniques are also compared. Conclusions drawn from this work are presented in Section VI.

ISBN: 978-1-61208-326-1

II. OVERVIEW: SITE NEIGHBOURHOOD ANALYSIS AND SITE RANKING

As discussed in Section I, the GIS-based site selection process can result in a number of potential sites or a relatively large potential area that meets the basic criteria, set in the first stage of selection process. In the second stage of site selection, it is important to reduce and prioritise the equal potential sites identified for further investigation and final selection. Some of the spatial multi criteria decision analysis techniques described earlier, may also involve the user's judgments and preferences over the relative importance of key indicators. Site neighbourhood comparison and ranking can be useful in a logical ordering of the available options based on only the analysis of the key indicators in the site neighbourhood. Once key indicators are defined and the effective neighbouring region is selected around the sites identified in the first stage of selection process, an appropriate ranking method can then be used to prioritise the alternatives.

Site neighbourhood analysis has been adopted in identifying potentially suitable sites for storm water harvesting in an urban area [5] in which a similar two-stage approach for site selection and ranking was adopted. The application of the site neighbourhood analysis has also been reported in selection of temporary municipal storage waste sites in Sweden using buffer analysis [6]. In the mentioned application, key demographic metrological indicators were analysed in these buffers around the sites [6].

For ranking the alternates or equal potential sites, different ranking techniques can be applied. Some commonly used ranking techniques are: Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [7], Superiority and Inferiority Ranking (SIR) [8] and Weighted Linear Combination (WLC) [9]. TOPSIS is among widely adopted methods for ranking. In the toolkit presented in this paper, TOPSIS method has been adopted along with a new approach which is based on a criterion sorting mechanism.

III. DEVELOPMENT OF THE TOOLKIT

As described previously, a site neighbourhood comparison approach has been adopted after a first phase site selection process has been completed. Using the appropriate logic in the toolkit, suitable areas are narrowed down into the most suitable site, considering additional criteria. The following steps are adopted in the toolkit to analyse the information and generate the results:

- First, the user provides two GIS layers to the toolkit and related data are loaded in the memory. One layer holds all potential sites whereas the other layer presents information of the neighbourhood of each site for the analysis. It is noted that this information could vary from problem to problem, e.g., socio-economic indicators or environmental indicators in the surrounding regions.
- The toolkit then scales all the indicators between 0-1(where 0 is the minimum value and 1 is the maximum value of each indicator in the layer). During the scaling process, i.e., commensuration, the user defines whether a particular indicator is Benefit (the more, the better) or Cost (the less, the better) in nature. The user can use either original values or scaled values. For scaling, the toolkit provides Maximum Score Procedure and Score Range Procedure, as in [10] and [11]:
  a. Maximum Score Procedure:

$$\text{(Benefit)}\quad X'_{ij} = \frac{X_{ij}}{X_{j.max}} \tag{1}$$

$$\text{(Cost)}\quad X'_{ij} = 1 - \frac{X_{ij}}{X_{j.max}} \tag{2}$$

  b. Score Range Procedure:

$$\text{(Benefit)}\quad X'_{ij} = \frac{X_{ij} - X_{j.min}}{X_{j.max} - X_{j.min}} \tag{3}$$

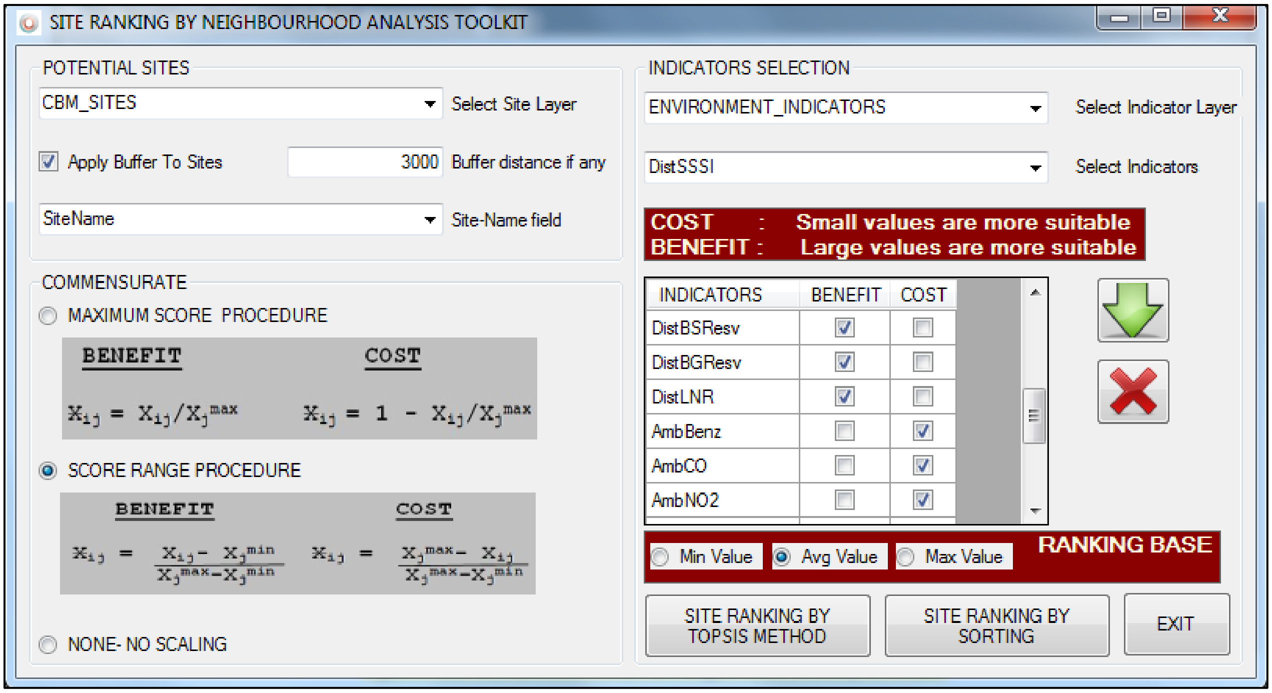
$$\text{(Cost)}\quad X'_{ij} = \frac{X_{j.max} - X_{ij}}{X_{j.max} - X_{j.min}} \tag{4}$$

  where X_{ij} is the value of the i^{th} location (potential site) for the j^{th} criterion. X'_{ij} is the scaled (standardized) value of X_{ij}. $X_{j.min}$ and $X_{j.max}$ are the minimum and maximum values of the j^{th} variable in the entire dataset [11].

  c. Using original values: This is the case when all indicators have same unit of measurement and they are either cost or benefit in nature. In such scenarios, decision maker can keep them in original units.
- The toolkit then selects the neighbouring areas of each potential site based on criteria specified by the user. This is done by applying buffers around the candidate sites and selecting the intersecting regions in indicator layer.
- The toolkit calculates the minimum, maximum and average values of each indicator in the selected neighbourhood of each site. Either of these average, maximum or minimum values can be selected to rank the sites.
- The final step is to assign ranks to the sites. User can choose the ranking either based on CSM or the TOPSIS method. The toolkit generates the results and visualises them in decision assistive graphs and tables.

The toolkit has been developed using Microsoft .Net C# programming language and an open source .Net spatial library, i.e., DotSpatial [12]. The toolkit uses Shapefiles, which is a "de facto" data type standard for vector data in GIS. The toolkit can also work on layers from an open source spatial database, namely SpatiaLite [13]. The problem is presented to the toolkit through two information layers (Shapefiles or SpatiaLite layers).

Figure1 presents the Geographical User Interface (GUI) of the toolkit. User can provide the necessary information using this interface. The user can also assign the buffer radius in map units to define the surrounding areas of each site to be included in the analysis. The toolkit first generates the buffer polygons around each site according to the user defined buffer size. Then the second layer containing indicators is intersected with these buffer polygons. If the buffer option is unused, only those areas are selected from the indicator layer that directly intersects with the sites without using any buffer. The toolkit can rank the sites based on the average, maximum or minimum value of each indicator in the given surrounding regions of each site.


SITE RANKING BY NEIGHBOURHOOD ANALYSIS TOOLKIT
SITE RANKING BY TOPSIS METHOD
SITE RANKING BY SORTING
EXIT
RANKING BASE

POTENTIAL SITES

Select Site Layer: CBM_SITES

☑ Apply Buffer To Sites — Buffer distance if any: 3000

Site-Name field: SiteName

COMMENSURATE

○ MAXIMUM SCORE PROCEDURE

BENEFIT: $x_{ij} = X_{ij}/X_j^{max}$ COST: $x_{ij} = 1 - X_{ij}/X_j^{max}$

● SCORE RANGE PROCEDURE

BENEFIT: $x_{ij} = \frac{X_{ij} - X_j^{min}}{X_j^{max} - X_j^{min}}$ COST: $x_{ij} = \frac{X_j^{max} - X_{ij}}{X_j^{max} - X_j^{min}}$

○ NONE- NO SCALING

INDICATORS SELECTION

Select Indicator Layer: ENVIRONMENT_INDICATORS

Select Indicators: DistSSSI

COST : Small values are more suitable
BENEFIT : Large values are more suitable

INDICATORS	BENEFIT	COST
DistBSResv	☑	☐
DistBGResv	☑	☐
DistLNR	☑	☐
AmbBenz	☐	☑
AmbCO	☐	☑
AmbNO2	☐	☑

○ Min Value ● Avg Value ○ Max Value

Figure 1.The user interface of the site neighbourhood analysis toolkit.

Each site may obtain different ranks for different indicators. A cumulative rank is then constructed by the toolkit using a Criterion Sorting Mechanism (CSM). If CSM is used, this cumulative rank is based on the individual ranks for each indicator. For this purpose, a rank sum is constructed for each site by summing up individual ranks of all the indicators. Sites are sorted in ascending order in terms of this rank sum. The site with lowest rank sum gets the overall Rank 1.

In order to compare the results of CSM technique, TOPSIS method is also incorporated within the toolkit. TOPSIS is selected because it is a commonly used ranking method in MCDA problems [14] [15]. It ranks the sites based on their distances from the most ideal and the least ideal solution [7].

If TOPSIS method is used, the cumulative site ranks are constructed using the empirical formulation of TOPSIS method. The procedure adopted in TOPSIS approach follows the steps provided in [7]. The detail of the calculation and procedure is therefore not provided here. The closer the rank is to 1, the priority of the alternate is higher as it is closer to the ideal solution and far from the worst solution [7]. It is a matter of sorting all the alternatives on the basis of these values and then assigning ordered ranks between 1-m, where 1 is the highest rank and m is the lowest rank.

The results are generated in the form of a report containing charts and a table, which provide the ranks of each site with respect to the indicators and also present the cumulative rank produced by CSM and TOPSIS. For charting, Microsoft chart control is used under the Microsoft Public Licence (Ms-PL) [16]. Polar chart scheme is also used to show the area covered by each potential site over different axis (indicators).

IV. Application Case Study

An application of the toolkit for the site neighbourhood analysis of six potential sites for CBM recovery in Wales, UK is presented. Using CBM technology, the gas contained in deep and un-minable coal seams can be exploited [17]. The gas recovery can also be enhanced by injecting carbon dioxide into coal seams (ECBM) [17].

A number of potential areas for CBM have been identified and reported in the North and South Wales coalfields [18]. Estimations have been carried out based on parameters, such as coal seam thickness, clean coal thickness (total thickness minus 15% ash and dirt allowance), density and gas content [18].

Two areas in Wales have been selected, one in the South coalfield (CBM area 4) and one in the North coalfield (CBM area 3). The areas have been selected among the potential CBM areas identified and reported in [18] based on “resource” capacity consideration. It is noted that the selected regions are not necessarily the most suitable areas for CBM in Wales. However, according to the CBM resource assessment figures given in [18], these areas may also contain a considerable CBM resource potential. The major characteristics of these two CBM areas are shown in Table 1.

Six 500m×500m squared areas have randomly been selected within the two regions mentioned (3 in each region). Figure 2 presents the coal resources in Wales and the potential CBM sites selected for this study. Since the area and properties of all the six sites are similar, it was assumed that the sites are equal potential candidates for CBM.

TABLE 1. ANALOGOUS CBM AREAS IN NORTH AND SOUTH COALFIELDS [14].

Properties	South Wales (Area 4)	North Wales (Area 3)
Coal thickness (m^3/t)	23.8	23
Clean coal thickness (m^3/t)	20.23	19.55
Coal density (g/cm^3)	1.33	1.26
Gas content Ave (m^3/t)	8.5	8

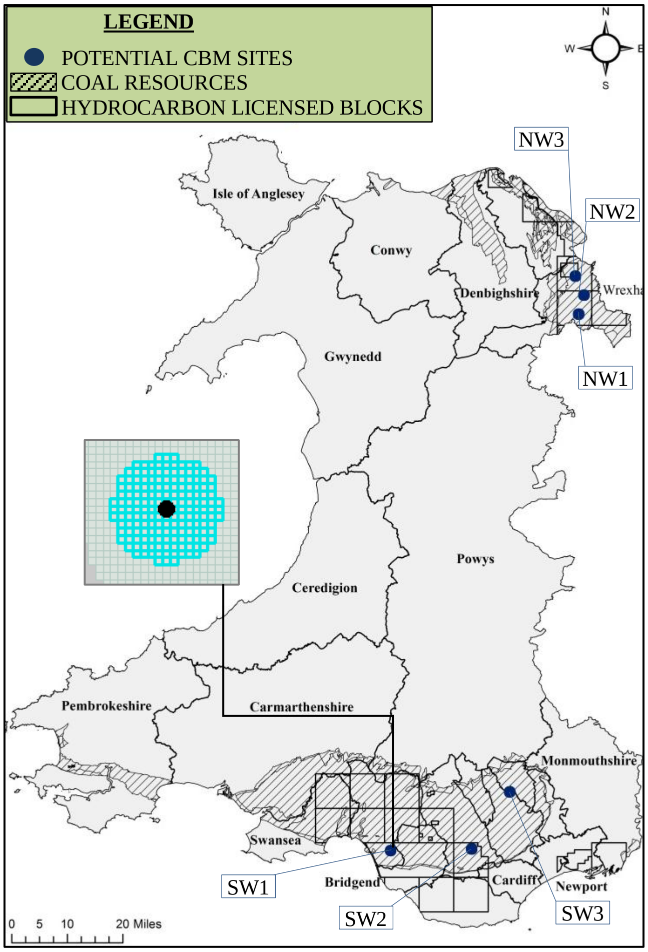


Figure 2. Coal resource map in Wales and selected potential CBM sites with 3 km buffers.

In the example provided, multiple sites exist with similar suitability, which is the same as the site suitability problem discussed in section two. The site neighbourhood analysis tool has therefore been applied, using some key environmental indicators. Some of these indicators were taken from the Department for Environment Food and Rural Affairs (DEFRA). The indicators are the background air pollution maps used for the ambient air quality assessment [19], which have also been adopted in the construction of the environmental deprivation within the Welsh Index of Multiple Deprivation 2011 (WIMD) [20]. Other environmental indicators used are the distance of each potential site from the protected areas in Wales, which was taken from datasets managed by the Countryside Council for Wales (CCW) [21].

The Intrinsic Evaluation Matrix of LandMap (Visual Sensory) dataset developed by CCW was also used for aesthetic coverage across Wales. This data set contains records of the ordinary and spectacular landscapes and information about the physical, ecological, visual, historical and cultural landscape of Wales [21]. The Intrinsic Evaluation Matrix covers scenic quality, integrity, character and rarity of the area and an “overall” index, which divides the Welsh landscape into Outstanding, High, Moderate and Low values [21]. These qualitative values were converted to numerical values of 1.0, 0.75, 0.50 and 0.25 respectively, to be used in the analysis. Table 2 shows the datasets used in neighbourhood assessment of the six potential sites for CBM exploitation.

All these environmental indicators were combined together into one GIS layer. For this a grid of 500m×500m area was generated across Wales using ArcGIS software. Using “near” and “spatial join” tools of the ArcGIS, these cells were then populated with the indicators data. The six potential sites were then saved in another vector layer. The site neighbourhood tool was applied with a buffer of 3 km around the sites (as shown in Figure 2) and ranking of the sites was carried out using both CSM and TOPSIS methods.

In the TOPSIS method used in this study, a relative weight of the criterion is also provided by the user to emphasize the importance of one over the other [7]. In simple cases, the weights can be directly applied and in cases, where uncertainty exists about the individual weights, pairwise comparison method is implied to find the relative weights [22]. In both cases the sum of all the individual weights should be equal to 1. To simplify the procedure in the example provided, it is assumed that the selected environmental indicators are equally important in the CBM site ranking process therefore an equal weight is assigned to every indicator for the construction of the TOPSIS ranks.

V. RESULTS AND DISCUSSION

The results of the analysis are presented in Table 2. The ranking analysis has been carried out on the basis of the average values of each indicator in the given neighbourhood (3km in this case). The overall rank for the six potential sites was constructed using both CSM and the TOPSIS methods. In CSM, final ranks are constructed by summing up the individual ranks for each site and then sorting in ascending order. The site with the lowest rank sum was assigned the overall rank 1 whereas the site with highest sum was assigned rank 6 accordingly.

The TOPSIS method calculates a distance of each potential site (alternate) from the most ideal and least ideal solution and the final ranks are constructed using (5).

Figure 3 shows the neighbourhood analysis results for each site considering key environmental indicators. The area covered under each polar graph is reflected in the ranks produced by the tool. The scale on the polar graphs is shown between 0-1 as the score range procedure, defined by (3) and (4), was used to scale the indicators.

As described previously, some of the indicators are *"cost"* in nature, such as the air pollution or the aesthetic maps whereas others are *"benefit"* in nature, such as the distance from protected areas. By using the appropriate equations for scaling the data, it was ensured that all *"cost"* and *"benefit"* indicators were scaled between 0-1, where 0 is the least favourable and 1 is the most favourable value.

The results indicate that by using the CSM in ranking procedure, the site SW3 is overall the most suitable site in terms of the indicators used in the analysis (TABLE 2). However, using TOPSIS method, the site SW3 has been ranked as the third most favourable site. The site SW3 is ranked as the best site based on only two individual indicators, i.e., Distance from Ramsar Sites (DistRamsar) and PM10 (AmbPM10). It has been ranked as the least suitable site based on only one indicator, i.e., DistLNR. The TOPSIS takes into account the distance of the site from the most and least ideal solutions. Therefore the SW3 has not been ranked as the most suitable site using TOPSIS.

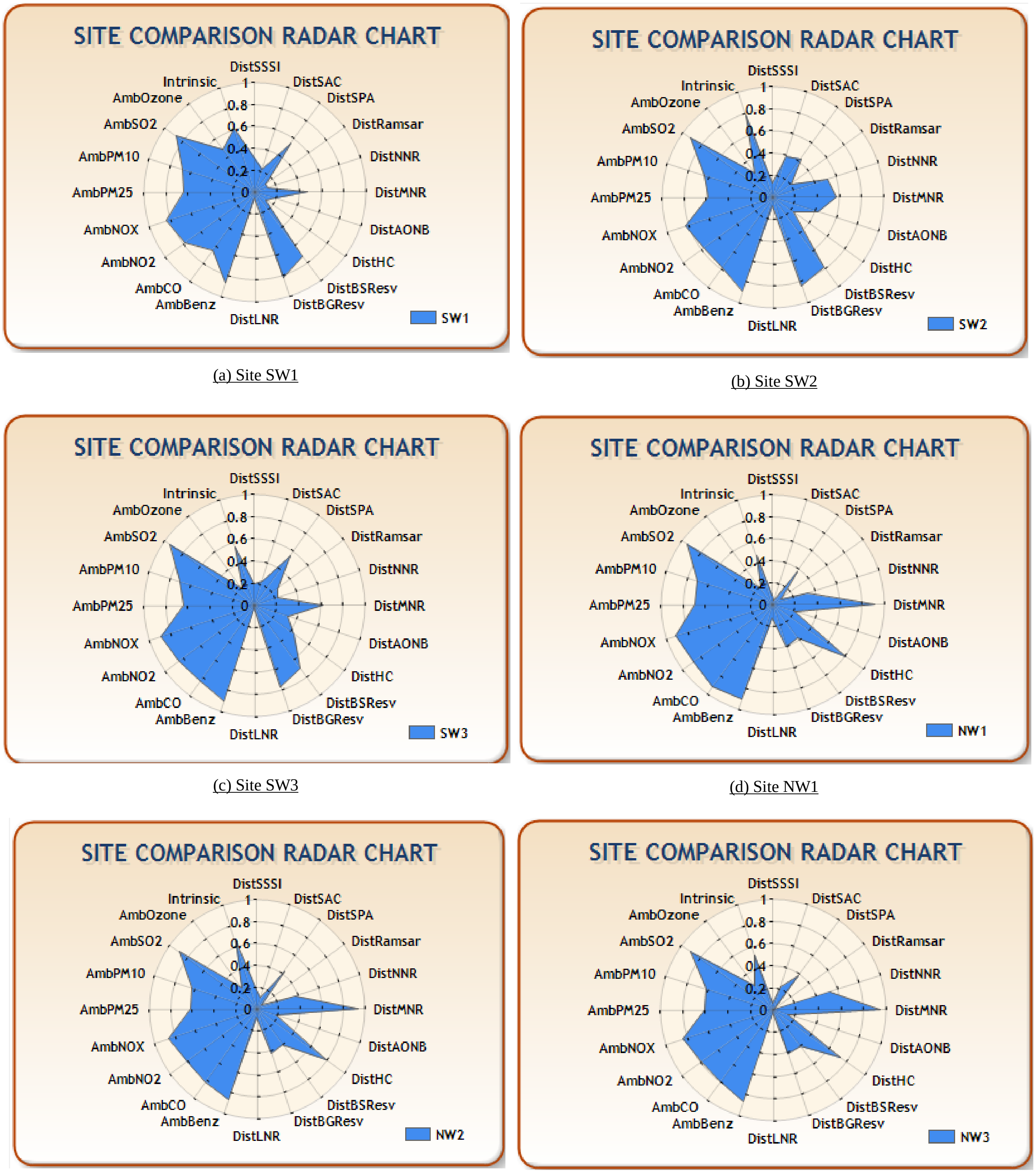
The site NW1 has been ranked as the most suitable site by TOPSIS and second most suitable site by the CSM technique. Exploring the individual ranks assigned by the CSM, it can be observed that the site NW1 has been ranked as the most suitable site based on eight different individual indicators and as the least suitable site based on seven individual indicators. As a result, this makes the NW1 less suitable site compared to the SW3 using the CSM. From the results obtained, it can be deduced that the outcomes from CSM technique are less affected by the extreme values of criterion, compared to the results obtained from TOPSIS.

The TOPSIS method assigns an overall Rank 2 to the site NW2 without it having top ranks for any of the individual indicators. Exploring the individual ranks further, it can be observed that almost all of the indicators ranks are in the middle of ranking scheme (1-6), i.e., taking the mean values. Considering the less deviated conditions of all indicators from the average values, the TOPSIS has ranked the site as the second most suitable alternative.

Both SW3 and NW1 are in the top three most suitable sites by using any of the two ranking methods described. The NW1 can overall be considered as the most suitable site with more confidence as it was rated as the top two sites by both ranking methods. Further confidence has been achieved by using the two methods of ranking together for the identification of the most suitable site, described in this application. It is noted that, once ranks are assigned to the sites, top ranked sites can be further investigated for environmental, health and socio-economic risks.

TABLE 2. ENVIRONMENTAL INDICATORS USED AND SITE RANKING RESULTS OBTAINED.

	INDICATORS			RANKS					
		Units	Abbreviation	SW1	SW2	SW3	NW1	NW2	NW3
1	Distance from SSI (Site of Special Scientific Interest)	m	DistSSSI	1	5	2	6	3	4
2	Distance from SAC (Special Areas of Conservation)	m	DistSAC	3	1	2	6	5	4
3	Distance from SPA (Special Protection Areas)	m	DistSPA	1	3	2	6	4	5
4	Distance from Ramsar Sites (Wetlands of International Importance)	m	DistRamsar	3	2	1	4	5	6
5	Distance from NNR (National Nature Reserves)	m	DistNNR	6	2	5	4	3	1
6	Distance from MNR (Marine Nature Reserves)	m	DistMNR	6	5	4	3	2	1
7	Distance from AONB (Areas of Outstanding Natural Beauty)	m	DistAONB	4	1	2	3	5	6
8	Distance from Heritage Coasts	m	DistHC	6	5	4	1	2	3
9	Distance from Biospheric Reserves	m	DistBSResv	2	1	3	6	5	4
10	Distance from - Biogenetic Reserves	m	DistBGResv	2	1	3	6	4	5
11	Distance from LNR(Local Nature Reserves)	m	DistLNR	4	3	6	1	2	5
12	Visual and Sensory Landscape overall Evaluation	-	Intrinsic	4	1	5	6	3	2
13	CO	mgm^{-3}	AmbCO	6	5	2	1	3	4
14	Benzene	μgm^{-3}	AmbBenz	5	3	2	1	4	6
15	NO_2	μgm^{-3}	AmbNO2	5	6	2	1	3	4
16	NOX	μgm^{-3}	AmbNOX	5	6	2	1	3	4
17	PM2.5(Particulate matter < 2.5 microns)	μgm^{-3}	AmbPM25	2	6	3	1	4	5
18	PM10 (Particulate matter < 10 microns)	μgm^{-3}	AmbPM10	3	4	1	2	5	6
19	SO_2	μgm^{-3}	AmbSO2	6	3	2	1	4	5
20	Ozone	μgm^{-3}	AmbOzone	1	3	5	6	4	2
	SITE RANK (CSM)			**5**	**3**	**1**	**2**	**4**	**6**
	SITE RANK (TOPSIS)			**6**	**5**	**3**	**1**	**2**	**4**



(a) Site SW1

(b) Site SW2

(c) Site SW3

(d) Site NW1

(e) Site NW2

(f) Site NW3

Figure 3. The results of the Site neighbourhood analysis for sites selected in this study.

VI. Conclusion

A GIS toolkit to analyse the neighbouring areas surrounding a potential site has been developed, which can be used to facilitate the process of decision making. The tool developed, provides a simple yet effective approach to deal with the ranking of sites based on some key indicators in their surrounding areas. Toolkit provides two different methods of site ranking, i.e., a simple Criterion Sorting Mechanism and the TOPSIS ranking method. This is useful where a number of equal potential sites or a large suitable area is acquired as the result of a GIS based site selection process in first phase of evaluation.

An application of the toolkit for CBM resource exploitation was presented. Six sites were randomly selected in the South and North Wales coalfields with similar resource potential. Based on selected environmental indicators, the potential sites were ranked, considering neighbourhood conditions. The results show the ranks for each site with respect to individual environmental indicators using both the methods.

The toolkit presented is part of a Spatial Decision Support System, developed to support a wide spectrum of geoenvironmental and geoenergy applications, where multiple criteria, such as the environment, public health, socio-economic and technical indicators are of importance in the decision process.

Acknowledgment

The work described in this paper has been carried out as part of the GRC's Seren project [23], which is funded by the Welsh European Funding Office (WEFO). Financial support for the first author is gratefully acknowledged.

References

[1] O. E. Demesouka, A.P. Vavatsikos, and K.P. Anagnostopoulos, Suitability analysis for siting MSW landfills and its multicriteria spatial decision support system: Method, implementation and case study. Waste Management, 2013. 33(5): pp. 1190-1206.

[2] P. V. Gorsevski, S.C. Cathcart, G. Mirzaei, M.M. Jamali, X. Ye, and E. Gomezdelcampo, A group-based spatial decision support system for wind farm site selection in Northwest Ohio. Energy Policy, 2013. 55(0): pp. 374-385.

[3] J. S. Jeong, L. García-Moruno, and J. Hernández-Blanco, A site planning approach for rural buildings into a landscape using a spatial multi-criteria decision analysis methodology. Land Use Policy, 2013. 32(0): pp. 108-118.

[4] M. A. Rahman, B. Rusteberg, R.C. Gogu, J.P. Lobo Ferreira, and M. Sauter, A new spatial multi-criteria decision support tool for site selection for implementation of managed aquifer recharge. Journal of Environmental Management, 2012. 99(0): pp. 61-75.

[5] P. M. Inamdara, S. Cookb, A. K. Sharmab, N. Corbyc, J. O'Connorc, and B. J. C. Pereraa, A GIS based screening tool for locating and ranking of suitable stormwater harvesting sites in urban areas. Journal of Environmental Management, 2013. 128(0): pp. 363-370.

[6] M. A. Ibrahim, G. Göransson, F. Kaczala, W. Hogland, and M. Marques, Characterization of municipal solid waste temporary storage sites: Risks posed to surrounding areas as a consequence of fire incidents. Waste Management, 2013. 33(11): pp. 2296-2306.

[7] C. L. Hwang, and K. Yoon, Multiple attribute decision making, Methods and applications. 1981, Heidelberg: Springer-Verlag.

[8] A. Rebai, BBTOPSIS: a bag based technique for order preference by similarity to idealsolution. Fuzzy Sets and Systems, 1993. 60(2): pp. 143-162.

[9] J. Malczewski, On the Use of Weighted Linear Combination Method in GIS: Common and Best Practice Approaches. Transactions in GIS, 2000. 4(1): pp. 5–22

[10] B. H. Massam, Multi-Criteria Decision Making (MCDM) techniques in planning. Progress in Planning, 1988. 30, Part 1(0): pp. 33.

[11] J. Malczewski, GIS and Multicriteria Decision Analysis. Wiley. New York, pp. 116-117, 1999.

[12] DotSpatial Library. [online] Available from: http://dotspatial.codeplex.com 2013.11.13

[13] SpatiaLite Library. [online] Available from: https://www.gaia-gis.it/fossil/libspatialite/home 2013.11.13

[14] Y. Chen, K.W. Li, and S. f. Liu, An OWA-TOPSIS method for multiple criteria decision analysis. Expert Systems with Applications, 2011. 38(5): pp. 5205-5211.

[15] J. Jia, Y. Fan, and X. Guo, The low carbon development (LCD) levels' evaluation of the world's 47 countries (areas) by combining the FAHP with the TOPSIS method. Expert Systems with Applications, 2012. 39(7): pp. 6628-6640.

[16] Microsoft .Net Chart Control. [online] Available from: http://archive.msdn.microsoft.com/mschart 2013.11.13

[17] C. M. White, et al., Sequestration of carbon dioxide in coal with enhanced coalbed methane recovery - A review. Energy and Fuels, 2005. 19(3): pp. 659-724.

[18] N. S. Jones, et al., UK Coal recourses for new exploitation technologies. 2004.

[19] DEFRA Modelled background pollution maps. [online] Available from: http://uk-air.defra.gov.uk/data/pcm-data 2013.11.13

[20] Welsh Government, Welsh Index of Multiple Deprivation. [online] Available from: http://wales.gov.uk/topics/statistics/theme/wimd/?lang=en 2013.11.13

[21] Countryside Council for Wales. [online] Available from: http://www.ccw.gov.uk/default.aspx 2013.11.13

[22] J. Malczewski, GIS and Multicriteria Decision Analysis. Wiley. New York, pp. 182-187, 1999.

[23] GRC SEREN Project. [online] Available from: http://grc.engineering.cf.ac.uk/research/seren/ 2013.11.13

Conflation in Geoprocessing Framework – Case Studies

Recent development and looking ahead

Dan Lee, Weiping Yang, and Nobbir Ahmed
Geoprocessing
Esri, Inc
Redlands, USA
e-mail: dlee@esri.com; wyang@esri.com; nahmed@esri.com

***Abstract*—Multiple sources of geographic information systems (GIS) data have been more easily and frequently produced and updated than ever. Yet, the traditional problem of data inconsistency spatially and in attribution, as the result of different ways of collecting and modeling data over time, remains obstacles in using the data for analysis and mapping. Efficient tools for data conflation have become a necessity for GIS users. Our recent work on conflation tools for the 10.2.1 desktop release of ArcGIS (the commercial GIS software by Esri Inc.) focused on linear feature matching techniques for identifying matching and no-match features. The initial results have proven time-saving in reconciling data for better positional and attribute quality and harmonization. Future challenges lay in formalizing data preparation, handling other feature types, and optimizing feature matching and workflows.**

Keywords-conflation; geoprocessing, feature matching; change detection; spatial adjustment; attribute transfer; workflow.

I. INTRODUCTION

GIS data maintained by many organizations and government agencies or obtained from data providers often need to be used together for multiple purposes of analysis and mapping. However, you may find that when displaying spatially overlapping or adjacent data, features representing the same ground locations or objects don’t line up even in the same map projection; or that a spatially up-to-date data lacks the desired attributes that only exist in another data source. Conflation is the process of matching corresponding features and making spatial adjustment or attribute transfer between them to improve data quality and consistency.

Within the geoprocessing framework, six new tools, shown in Fig. 1, have been developed for the 10.2.1 desktop release of ArcGIS. This development is an advance from the legacy technology used in Esri’s earlier product [1].

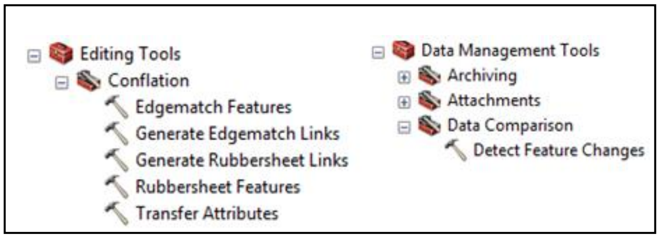


Figure 1. Conflation tools (by the tool icon 🔨) inside Editing and Data Management toolboxes in ArcGIS.

Good conflation outcome is achievable as a result of high feature matching accuracy; inspection and editing may be necessary as part of the workflows. The automation and the significantly reduced manual work enable GIS users to move away from living with imperfect data and to reach higher standards in geographic data integration, analysis, and mapping more efficiently.

This paper is organized as follows: Section II briefly reviews the feature matching processes and associated tools. Our initial efforts have focused on linear features. Section III presents a few conflation scenarios and workflows used to accomplish the tasks with success and efficiency. Section IV gives conclusions and thoughts on future work.

II. FEATURE MATCHING IN CONFLATION TOOLS

At the core of conflation is feature matching, either between overlapping datasets or between adjacent datasets. Feature matching accuracy relies highly on data quality, similarity, and complexity. The feature matching techniques used in the conflation tools are briefly described below.

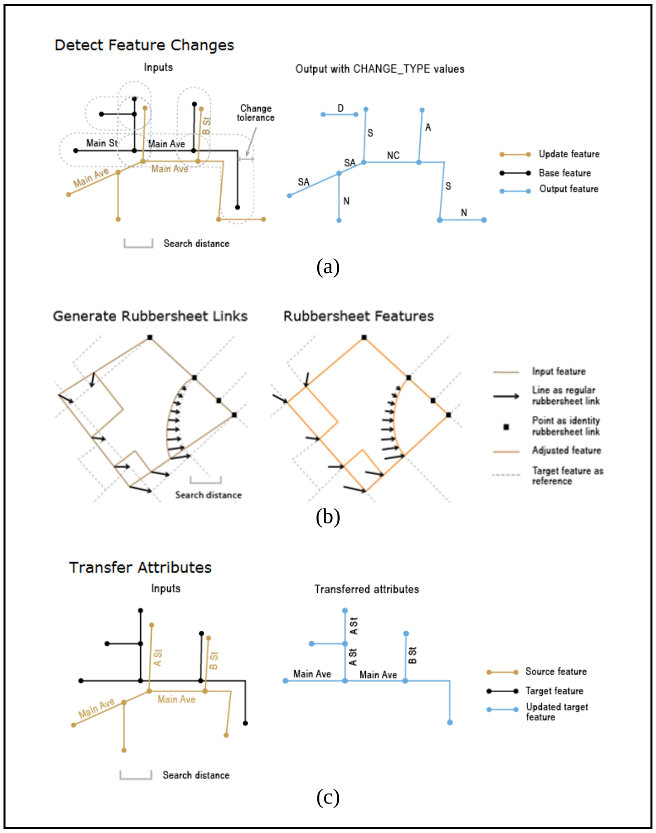
A. *Feature matching of overlapping datasets*

There have been many research papers and implementations on feature matching of overlapping datasets. Some examples include: a five-step statistical approach with the use of a merit function to compute unique combinations of matching pairs among potential but ambiguous matching pairs [2]; a delimited stroke oriented algorithm consisting of four processes for the matching of road networks [3], and an optimization model for linear feature matching which takes into account all potential matched pairs simultaneously by maximizing the total similarity of all matched features [4].

The feature matching technique we choose to use is based on the fundamental analysis of the topological structures and feature pattern recognition. The key processes are: (1) analyzing feature topology, i.e., to find nodes and joining lines in linear features, (2) building structures (paths and patterns), (3) matching structures, and (4) matching features within structures. More details on this feature matching approach are given in a separate paper [5].

The matching information can be written out to a match table with five fields: SRC_FID (source feature ID), TGT_FID (target feature ID), FM_GRP (feature match group ID), FM_MN (matching relationship in the form of m:n, where m and n represent the numbers of source features and target features respectively in a match group and can be greater or equal to 1), and FM_CONF (feature matching confidence level with values between 0 and 100). This feature matching process is the basis for the following tools which help perform conflation tasks on overlapping datasets that cover the same geographic areas:

- Detect Feature Changes (DFC) identifies spatial and attributes changes between update and base features. The output change types include: S for spatial change, A for attribute change, SA for spatial and attribute changes, NC for no change, N for new update feature, and D for potentially to-be-deleted base feature. See the illustration in Fig. 2–(a).
- Generate Rubbersheet Links (GRL) generates rubbersheet links, including regular links (lines going from matching source to target locations) and identity links (points where source and target locations are identical and not being moved in rubbersheeting adjustment). The tool Rubbersheet Features (RF) does rubbersheeting adjustment using the generated links to align source features with target. See the illustration in Fig. 2–(b).
- Transfer Attributes (TA) transfers feature attributes from source to matching target features. See the illustration in Fig. 2–(c). By design when multiple source features match one or more target features, attributes from the first picked source feature are transferred to all matched target features.



(a)

(b)

(c)

Figure 2. Feature matching based tools for conflation of overlapping datasets.

B. *Edgematching - matching features of adjacent data areas*

Edgematching is the process of identifying corresponding features along the edge (meeting locations) of adjacent (side-by-side) datasets. The key processes are: (1) finding features within the specified search distance to each other along their meeting areas, (2) evaluating the geometric characteristics and continuity from input to adjacent features and vice versa, and (3) determining the best fit pairs of corresponding features. The tools for edgematching are:

- Generate Edgematch Links (GEL), which generates edgematch links, followed by Edgematch Features (EF) to adjust features to new connecting locations guided by the links. See the illustrations in Fig. 3.

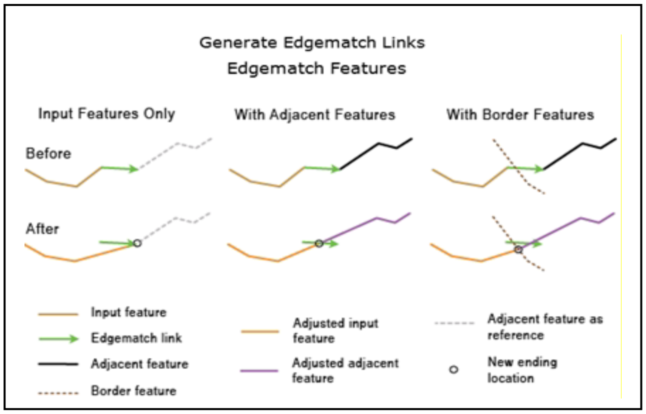


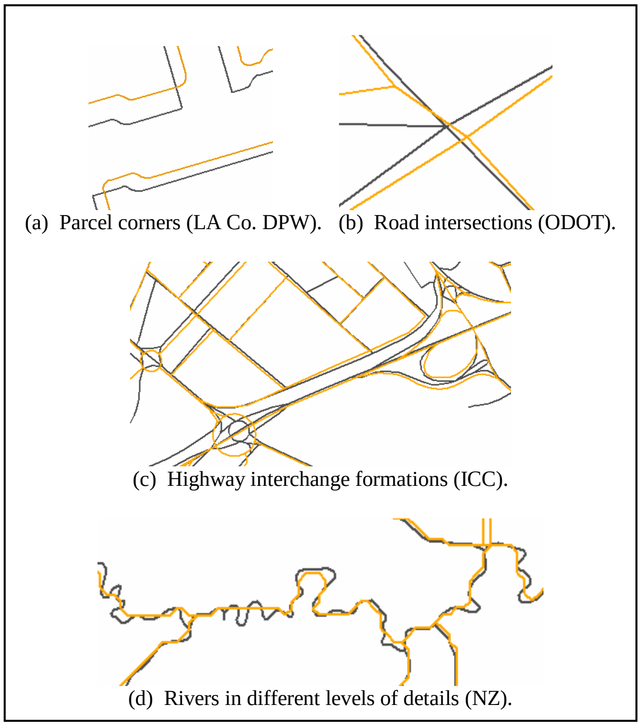
Figure 3. Edgematching by moving endpoints (one of the available options) of features to new connecting locations.

C. *Challenges in feature matching*

No matter how sophisticated the feature matching techniques are, the reality of geographic data is often more challenging than the automatic analysis can handle. The main factors causing difficulties and errors in feature matching include:

- Invalid feature topology, such as gaps, overshoots, undershoots, overlaps, and duplicates.
- Differences in feature representations and data modeling of the same ground objects, especially between overlapping data sources, ranging from variations in their geometric characteristics to distinctions in their structural formations. For example: round vs. squared corners of parcel boundary lines in Fig. 4–(a), one vs. separate road intersections in Fig. 4–(b), and different collections of road merging or splitting around complex highway interchange areas in Fig. 4–(c).
- Features with different levels of details for multiple map scales. See the illustrations in Fig. 4–(c) and (d).

The more dissimilar the corresponding features the harder to make the right feature matching decisions. The conflation tools can take into account common attributes between input datasets to help determine the right match, but the common attributes are often either unavailable or incomplete.



(a) Parcel corners (LA Co. DPW). (b) Road intersections (ODOT).

(c) Highway interchange formations (ICC).

(d) Rivers in different levels of details (NZ).

Figure 4. Examples of dissimilar overlapping features.

Given the highly automated conflation tools and the possibility of some mistakes in the results, the question we need to address is what it will take to find and correct errors and to complete the conflation tasks. That leads to the discussion on conflation workflows. Due to the length limitation of the paper, the discussion focuses on conflation of overlapping data sources.

III. CONFLATION WORKFLOWS

Conflation tasks may be as simple as to make spatial adjustment or attribute transfer from one data source to another for better positional accuracy and attribute consistency, or as comprehensive as to unify information from multiple data sources for the best combined result. In general a conflation workflow may consist of three components: (1) preprocessing to eliminate input data issues and to exclude irrelevant features from participating in the conflation process; (2) automated processes using the conflation tools to produce mostly correct results and conflation evaluation tools to derive information that help identify potential mismatched features and find locations that need attention; and (3) interactive review and editing based on the automatically derived information, as well as visual inspection, to improve the result to satisfaction.

A few workflows are examined below using real world data. But before getting into that, it is necessary to briefly explain the preprocessing and conflation evaluation tools mentioned above.

Preprocessing is common to all conflation tasks. A few generic guidelines and possible geoprocessing tools to use are given below and are not repeated for every workflow scenario:

- Fix invalid geometry (Repair Geometry tool)
- Validate feature topology (Topology Tools)
- Remove overshoots and undershoots (Trim Line and Extend Line tools)
- Delete unwanted duplicates (Delete Identical tool)
- Break unintended long-running features at intersections (Feature To Line tool)
- Exclude irrelevant features from participating in conflation processes. (Select By Attributes or Select By Location)

Other data specific preprocessing may also be necessary; it is important to identify data issues and use appropriate tools to resolve them.

The conflation evaluation tools mentioned in workflow component (2) above have been built either by Python scripting or by chaining together existing geoprocessing tools. They produce information to help understand the conflation results, identify potential errors, and facilitate the interactive review and editing processes. They are supplementary tools and do not come with the release. Here are the main evaluation tools:

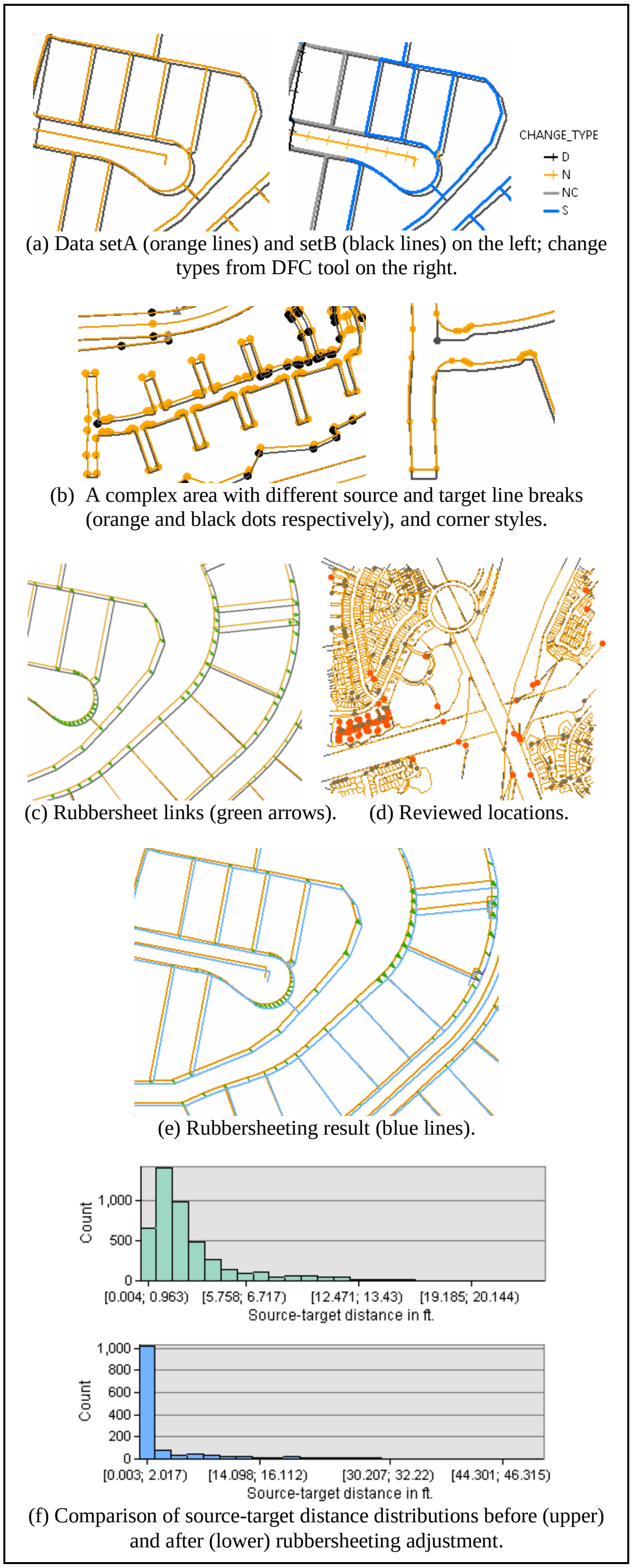
- Check Feature Matching (CFM) – analyzes the feature matching information produced by DFC, GRL, or TA tools and flags questionable matched conditions, for instance, the multiple source or target features in a m:n match group don’t belong to the same line or the matched source and target features may be too far apart to be the right match.
- DFC and Evaluation – runs DFC tool and checks for potential change type errors caused by mismatches; it is especially helpful to verify change types D and N so their source features can be excluded from participating in GRL and TA processes as needed. The tool also makes a bar graph for change types.
- GRL and Evaluation – runs the GRL tool and produces point features at locations where the generate rubbersheet links intersect or where no links are generated. It also adds source (from-point) and target (to-point) vertex types to the links. The vertex types are simply: 0 for in-line vertex, 1 for dangle end, 2 for pseudo node, 3 for T-node, 4 for cross-node, 5 for node with 5 joining lines, and so on. This information is intended to facilitate the inspection, especially at major intersections, for example if a link starts at type 4 and ends at 1, it may not be linking corresponding locations.
- RF and Assessment – It runs RF tool to perform rubbersheeting adjustment and produces additional data and information to help compare the source data and its adjusted result and to assess the location accuracy improvement.
- TA and Evaluation – It runs TA tool and produces additional data and information to help inspect no transfer and potentially mis-transferred cases.

Using real world test data, two conflation scenarios and workflows were examined: A. rubbersheeting spatial adjustment workflow; B. a more comprehensive workflow requiring both spatial and attribute unification.

A. Rubbersheeting spatial adjustment workflow

The data used to demonstrate this workflow are two sets of parcel lines (provided by LA Co. DPW); let's name them setA (3779 lines) and setB (3840 lines) as shown in the left image of Fig. 5-(a); preprocessing details are omitted here. The goal was to spatially adjust setA towards the more accurate setB. The workflow steps, actions, and results are:

- Step 1: Ran the DFC and Evaluation tool – see change types in the right image of Fig. 5-(a). Notice that both setA and setB contain lines that were not parcel lines and didn't have corresponding features. They ended up being N and D change types. Actions were taken to verify them and exclude them from GRL process in Step 2 for better result.
  - Through flagged information and visual inspection, 86% of the Ns (not matched in setA) were confirmed; others corrected. Most of the errors occurred in one large area with not only complex feature shapes but also a big contrast in the number of line breaks (orange dots for source line breaks; black dots for target line breaks) and corner styles, shown in Fig. 5-(b).
  - Through flagged information and visual inspection, 99% of the Ds (not matched in setB) were confirmed; others corrected.
  - Selected 3056 matched lines from setA and 2915 from setB, excluding the verified Ns and Ds respectively, as inputs for Step 2 below.
- Step 2: Ran the GRL and Evaluation tool - total 4413 regular rubbersheet links (see Fig. 5-(c) for a close up) and 0 identity link were generated.
  - Reviewed the flagged no link locations (red dots in Fig. 5-(d), mostly concentrated in the southwest area, i.e., the complex area shown in Fig. 5-(b)), and added 65 critical links.
  - Through flagged intersecting links (brown dots in Fig. 5-(d)) and other hints and inspections, total 104 links were modified and 29 deleted.
  - Analyzed the feature matching result; the estimated accuracy value breakdowns are presented in Table I. A 98.34% high accuracy was reached among matched features, while the overall accuracy 93.84% was largely affected by the no match cases in the complex area.
- Step 3: Ran the RF and Assessment tool – setA was adjusted, as shown in Fig. 5-(e), using total 4449 rubbersheet links. Among many possible ways of measuring positional alignment improvement, the following two are simple and effective:
  - Compared rubbersheet link counts from Step 2 and from rerun of GRL after rubbersheeting: regular link count reduced from 4413 to 1200; identity link count increased from 0 to 3102. This indicates over 70% of source locations are perfectly adjusted to target locations.
  - Compared source to target (source-target) distance distributions through the lengths of the regular links: the distances were obviously more concentrated in the shorter range after rubbersheeting adjustment; see Fig. 6-(f).



(a) Data setA (orange lines) and setB (black lines) on the left; change types from DFC tool on the right.

(b) A complex area with different source and target line breaks (orange and black dots respectively), and corner styles.

(c) Rubbersheet links (green arrows). (d) Reviewed locations.

(e) Rubbersheeting result (blue lines).

(f) Comparison of source-target distance distributions before (upper) and after (lower) rubbersheeting adjustment.

Figure 5. Rubbersheeting workflow (LA Co. DPW parcel data).

TABLE I. ESTIMATED FEATURE MATCHING ACCURACY IN STEP 2

Matched Feature Groups				
Match Relation -ship	***Group Count (Gc)***	***Correct group count (Cgc)***	***Error group count (Egc)***	***Accuracy (percentage of Cgc/Gc)***
1:1	2267	2244	7	98.99%
m:n	384	363	21	94.53%
Total	**2651**	**2607**	**28**	**98.34%**
Unmatched Features				
Match Relation -ship	***Feature count (Fc)***	***Correct feature count (Cfc)***	***Error feature count (Efc)***	***Accuracy (percentage of Cfc/Fc)***
1:0[a]	116	5	111	4.31%
0:1[b]	26	9	17	34.62%
Total	**142**	**14**	**128**	**9.86%**
Grand total	**2793**	**2621**	**156**	**93.84%**

a. Change type N features of DFC output. b. Change type D features in DFC output.

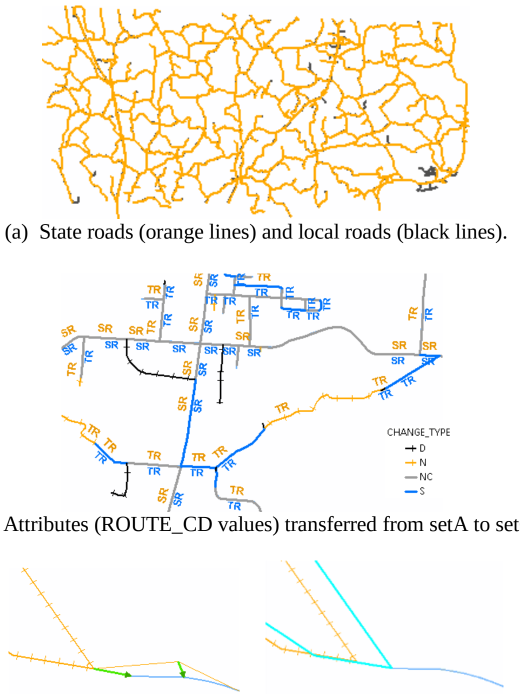
This test data was intentionally chosen for its challenging conditions including the seemingly CAD-imported features with no attributes to separate road centerlines and other features from the parcel lines and the inconsistent data modeling. Also on purpose, only minor preprocessing was done so the strengths and weaknesses of the conflation tools could be tested using near raw data. Although unaccounted errors may exist and would slightly lower the estimated accuracy levels, this exercise produced encouraging result.

B. *Workflow of unifying datasets for the best outcome*

The data used to demonstrate this workflow are two subsets of state and local roads in northeast area of Meigs County (provided by Ohio DOT) as shown in Fig. 6-(a). Let's name them setA (775 lines) and setB (827 lines), knowing that setB is spatially more up-to-date and accurate than setA. The goal was to produce a unified output with the spatial accuracy of setB, the uncommon attributes from both sets for matched features, and all unmatched features of both sets properly positioned keeping their original attributes.

There could be various ways to get there; all would be quite comprehensive. Below is one of the possible workflows attempted in this study. Good preprocessing was done to break lines where necessary, especially for route features in setA, and to improve the topological consistency between the two sets; details are omitted here. The conflation workflow strategy was to: (1) make a setC by copying setB so it has the spatial accuracy and attributes of setB intact, (2) transfer desired attributes from setA to setC for matched features, (3) identify unmatched features in setA, (4) spatially adjust the unmatched features of setA towards setC, and (5) merge the adjusted unmatched features of setA into setC. Here are the details:

- Step 1: Copied setB to setC and ran DFC and Evaluation tool – Actions were taken to verify N and D change types and to exclude them from TA process in Step 2 for better result.
  - Through flagged information and visual inspection, 58 of the 61 Ns (unmatched in setA) were confirmed, 3 corrected; 109 of the 117 Ds (unmatched in setC) were confirmed; 8 corrected.
  - Selected 717 matched lines from setA and 718 from setC, excluding the verified Ns and Ds respectively, as inputs for Step 2 below.
- Step 2: Ran TA and Evaluation tool – see the example result of attribute transfer (ROUT_CD) in Fig. 6-(b), which is superimposed with change types from Step 1. Actions were taken to verify TA result:
  - Reviewed the only 4 no transfer records in setC. Manual transfer was needed.
  - For each of the 39 m:n match cases, attributes from one of the m features were transferred to all n features in the match group by design. Review of the flagged cases and corrections would be needed, if the default transfers were undesired.
  - Analyzed the feature matching result; the estimated accuracy value breakdowns are presented in Table II. A 100% accuracy was reached among matched features, while the slightly lower overall accuracy 98.74% was mainly affected by the few mistakenly unmatched cases.
- Step 3: Ran GRL tool - total 12322 regular rubbersheet links and 19 identity links were generated.



(a) State roads (orange lines) and local roads (black lines).

(b) Attributes (ROUTE_CD values) transferred from setA to setC.

(c) Rubbersheet links & Ns of setA. (d) Adjusted Ns merged in setC.

Figure 6. Unifying multiple data sources for best outcome.

TABLE II. ESTIMATED FEATURE MATCHING ACCURACY IN STEP 2

Matched Feature Groups				
Match Relation -ship	***Group Count (Gc)***	***Correct group count (Cgc)***	***Error group count (Egc)***	***Accuracy (percentage of Cgc/Gc)***
1:1	656	656	0	100%
m:n	39	39	0	100%
Total	**695**	**695**	**0**	**100%**
Unmatched Features				
Match Relation -ship	***Feature count (Fc)***	***Correct feature count (Cfc)***	***Error feature count (Efc)***	***Accuracy (percentage of Cfc/Fc)***
1:0[a]	61	58	3	95.08%
0:1[b]	117	109	8	93.16%
Total	**178**	**167**	**11**	**93.83%**
Grand total	**873**	**862**	**11**	**98.74%**

a. Change type N features of DFC output. b. Change type D features in DFC output.

- Reviewed the links focusing on where the Ns of setA were to be adjusted to connect with features in setC, see the example in Fig. 6-(c). All 58 Ns had links to target locations.

- Step 4: Ran RF tool to adjust the N features using the generated rubbersheet links.
- Step 5: Append the adjusted Ns (highlighted) of setA onto setC (blue lines), as shown in Fig. 6-(d).

This test data was chosen for its clean topology and relatively high similarity. The test results indeed proved a close correlation between high quality starting data and accurate conflation result. Subsequently, very little manual work was needed in this case study.

IV. CONCLUSIONS AND THOUGHTS ON FUTURE WORK

Through the two case studies using real world data, the new conflation tools developed for ArcGIS were successfully tested and produced high quality results. The first case study proved that the very costly and nearly impossible task of generating thousands of rubbersheeting links one-by-one manually could be done mostly automatically with small amount of interactive editing. The second case study gave a good consensus on how the ultimate goal of conflation, i.e., the fusion of multiple source information for the best unified single outcome can be achieved efficiently.

The development of highly automated conflation tools for linear features is a major step forward in supporting the reconciliation of multiple data sources – an important process for data integration and data sharing demanded and embarked on by GIS and mapping agencies [6][7]. Our future efforts will focus on the following areas:

- Enhancements on feature matching with additional pattern recognitions and richer output information.
- New tools for other feature types (point and polygon) and contextual conflation.
- Integrated interactive conflation inspection and finishing environment.
- Streamlining of workflows by testing more real world scenarios and making conflation support tools available at ArcGIS Resources Center: http://resources.arcgis.com/en/home/.
- Investigations on harmonizing spatially related features, such as utility lines and other boundaries spatially associated with parcel lines.
- Extending the use of conflation tools in other areas, such as data quality checking, linking multi-scale geospatial databases and cartographic representations for incremental updating.
- Potential of using image and other information sources in feature matching [8].

ACKNOWLEDGMENT

Thanks to the many Geoprocessing team members for their support and to the following organizations for sharing their data for testing and creating illustrations in this paper:

Parcel datasets, sourced from the Department of Public Works (DPW), Los Angeles County, USA.

Road datasets, sourced from Institut Cartogràfic de Catalunya, Barcelona, Spain.

River datasets, sourced from National Institute for Water and Atmospheric Research (NIWA) and Land Information New Zealand (LINZ). Crown Copyright Reserved.

Meigs county road data, sourced from Ohio State Department of Transportation, USA.

Also thanks to Mark Hollingshead, Esri St. Louis Office, USA, for providing well-prepared Meigs road data.

REFERENCES

[1] A. E. Lupien and W. H. Moreland, “A general approach to map conflation”, Proceedings of AutoCarto 8, March 30 - April 2, 1987, Baltimore, Maryland, USA, pp. 630-639.

[2] V. Walter and D. Fritsch, “Matching spatial data sets: a statistical approach”, International Journal of Geographical Information and science, vol. 3(5, 1999), pp. 445-473.

[3] M. Zhang and L. Meng, “Delimited stroke oriented algorithm – working principle and implementation for the matching of road networks”, Journal of Geographic Information Sciences, vol. 14(1), June, 2008, pp. 44-53.

[4] L. Li and M. Goodchild, “An optimization model for linear feature matching in geographical data conflation”, International Journal of Image and Data Fusion, vol. 2(4), 2011, pp. 309-328.

[5] W. Yang, D. Lee, and N. Ahmed, “Pattern Based Feature Matching for Geospatial Data Conflation”, submitted to GEOProcessing, 2014, Barcelona, Spain.

[6] L. Stanislawski, C. Nelson, and M. Hamann, “Automated Conflation of Reach Data for the National Hydrography Dataset”, http://proceedings.esri.com/library/userconf/proc02/pap1207/p1207.htm, [retrieved: Nov. 26, 2013].

[7] Y. Li and C. Liu, “Spatial approaches for conflating GIS roadway datasets”, Sustainable Transportation Systems, 2012, pp. 290-298, doi:10.1061/9780784412299.0035.

[8] C. Chen, C. Knoblock, and C. Shahabi, “Automatically conflating road vector data with orthoimagery”, GeoInformatica, 10(4), 2006, pp. 495-530.

Pattern Based Feature Matching for Geospatial Data Conflation

Weiping Yang, Dan Lee, Nobbir Ahmed
Geoprocessing
Esri, Inc.
Redlands, USA
wyang@esri.com; dlee@esri.com; nahmed@esri.com

***Abstract*—This paper describes a linear feature matching method for overlapping data sources aimed at developing geospatial conflation tools. This method is based on identifying distinguished feature patterns, which are categorized as atomic patterns and composite patterns. It is hoped that these feature level patterns, once identified and modeled, will serve as a signature and as a helpful vehicle to enrich semantic meanings of a dataset. Features from overlapping sources will then be matched, by first matching large or complex structures and then breaking down to individual features of varying cardinalities. This feature matching method has been implemented as a core component for geoprocessing conflation tools to perform spatial adjustment, attribute transfer, and feature change detection. Applying these tools, in workflows, to real world data has produced promising results.**

Keywords-geospatial featture patterns; feature matching; geospatial data conflation; geoprocessing.

I. Introduction

Geospatial data conflation is one of the classic requirements in data integration where attributes from one data source need to be transferred, or geometries adjusted, based on feature correspondences of another data source. With the vastly increasing distributed geospatial data and sharing of these data over the web, high demands on conflation solutions have been seen in recent years. The concept of conflation has also been extended to detect feature changes between updated and existing datasets where new features emerged and previously existing features disappeared.

Due to differences in data capturing purposes, methods, scales, accuracy, or collecting time, corresponding geospatial features have discrepancies which may be spatial and non-spatial. Spatial discrepancies can be topological, geometrical, and metrical. A road feature matching two roads broken at a new T-intersection, for example, is a topological change; a circular cul-de-sac corresponding to a rectangular loop reveals a geometrical difference; and a short extension of a road from an intersection finding its peer prolonged shifted and slightly rotated causes altered metric measures. A non-spatial discrepancy occurs when the road name “Main St.” in one dataset is meant to be “Main Street” in another. Hence, finding relationships between corresponding features must be equipped with capabilities of distinguishing and handling these discrepancies. As a general consensus, the automation of conflating geospatial data requires a multitude of processes and possibly iterations with these processes, among which correctly matching features is central to the success of the automation.

Section II of the paper overviews some published feature matching methods, and the one we have developed for the purpose of creating a suite of conflation tools. It is followed by a discussion in Section III on identifying distinct structural shapes which are the basis of matching features. The matching process, starting from matching structures is presented in Section IV. The result of structural matching is then broken down to match individual features, which is discussed in Section V. The testing of the feature matching method in building conflation tools and the practical uses of the tools in workflow contexts are discussed in Section VI. The paper will end with conclusions and a brief discussion on future work.

II. Overview of feature matching methods

Given two overlapping sets of geospatial features named A and B, the problem of feature matching can be phrased as: for each feature in A, find its most likely counterpart in B if it exists. The catch here is the modifier “most likely” since it implies the application of cognitive knowledge and a conclusion can be elusive in complex configurations. Finding a matching method supported by objective measures has been a challenge to academicians and practicing professionals working on Geographic Information Systems (GIS) over the past 30 years.

Ever since an automated, interactive feature matching system via iterative rubber-sheeting was developed by the U.S. Census Bureau [1], researchers have been looking to use similarity measures against features to enhance matches between datasets. Various definitions of similarities and their measuring methods have been discussed, from considering geometric properties such as orientation, shape, length, etc. to combining themes and semantic attributes [2], [3]. At the operational level, analytical [2], [4], statistical [5], and linear programming models have been proposed [6], [7].

In what follows, a method of matching overlapping linear features is first outlined, leaving elaboration of details in subsequent sections. This method is in line with some of the published approaches [2], [4] that find similarities between corresponding features of different sources, based on topological, geometrical and metric properties of features. The main differences of the proposed method are that it emphasizes on discovering shape patterns from a single dataset first and then cross-referring these patterns to probe and to assemble correspondent patterns in another. The method regards geometric patterns built progressively from low level shapes as objects to compare for similarities.

The concept of identifying a hierarchy of pattern structures in a geospatial dataset has been inspired by a discussion on modeling patterns and structures in maps [8]. In cities and towns, design patterns are generally observed in urban and regional planning [9]. Researchers in GIScience have been detecting and describing street network patterns in geospatial databases out of a mass of seemingly chaotic data collections [10], [11]. It is believed that such a pattern system would reveal "meanings" of geospatial data for intelligent queries and applications.

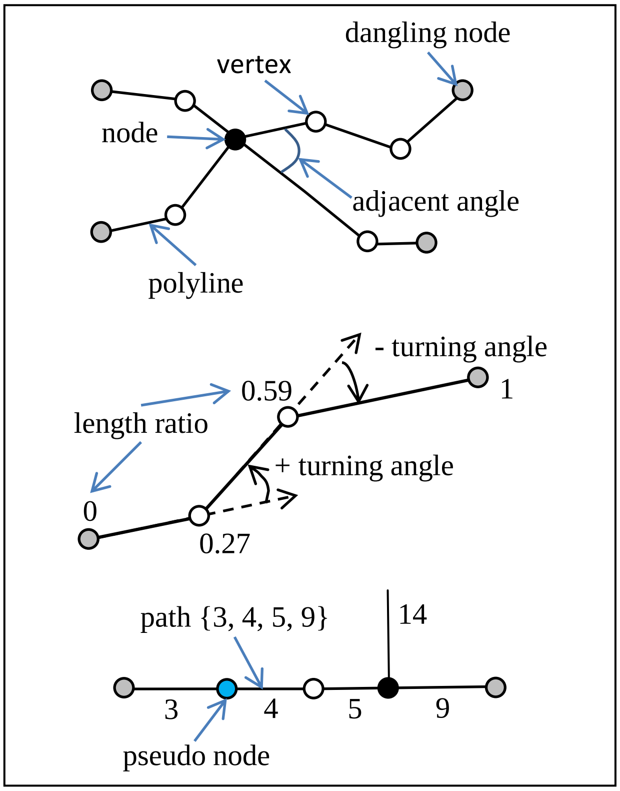


Figure 1. Illustration of terms.

Figure 1 illustrates the terms that are useful to the methods described in this paper. Linear features include roads, rivers, utility lines, or land use boundaries, etc. These features can be represented with polylines extended by a series of vertices in 2D space. A network of nodes and paths are formed from all input polylines. Informally, nodes are located at the two extreme vertices of a polyline. A node can be a connection of one or more polylines when at least one of their extremes coincide or are intersected at the same location. A dangling node does not make a connection to any other polylines while a pseudo node intersects exactly two polylines. A path is formed by one or more polylines connected at or incident to nodes and constrained by metric properties. In the method described in this paper, a path can span a number of nodes some of which may have more than two incident paths. Properties such as orientation, length, normalized length ratios of accumulative segments to the whole length, and turning angles are used to characterize a path. The number of adjacent paths and the adjacent angles formed by two adjacent paths are associated with a node. For clarity, the series of feature IDs composing a path is represented in a curled bracket as ordered list, as shown in Figure 1.

Individual polylines can be short and long, straight, curved, or otherwise forming various shapes. The matching method discussed here uses a combination of top-down and bottom-up approach. The *bottom-up process* constructs progressively larger structures by connecting polylines that are constrained to form descriptive shapes or patterns in one dataset. By obtaining well behaved large patterns such as highways or river networks, global information about the data can be grasped. Models and analysis on top of patterns could be developed to derive semantics about the data.

The *top-down process* takes each of the pattern structures to locate and to match similar patterns composed of a single or a plural set of constructive patterns from the other dataset. The matches of a few well distributed, distinct, and robust patterns could shed light on how the two datasets are shifted and rotated. The information could be further used for constraining proximity searches for matching of smaller or weakly-determined shapes. On the other hand, if the number of mismatches among the pattern structures is high, the matching process could report a strong dissimilarity between the two datasets. After the process of matching structures is completed, matching individual features will be followed by breaking down a pair of matched structures into corresponding feature pairs. In the break-down process, a method of gauging "affiliation" between features using metric and topological properties will be applied.

III. Identifying Structures in a Dataset

Two kinds of structures can be identified in a dataset alone without referring to an overlapping or context layer. The first kind, termed atomic patterns, is formed by a single polyline geometry of a feature. The second, termed composite patterns, consists of a series of polyline geometries of two or more features which themselves may be of atomic patterns. Once identified, all these patterns and supporting polylines can be organized with a dynamic hierarchical structure for easy manipulation.

A. *Atomic Patterns*

Atomic patterns are captured when shapes of a linear feature class are read in and cached. No searches are involved in this stage. Depending on the nature and purpose of the feature class, there are unlimited ways that a linear feature can be shaped. It is impossible to design all stereotypes to fit all features in an ordinary feature class. Nevertheless, certain shapes stand out to observers' eyes; others don't. It is possible to devise well-structured scalable stereotypes to filter and describe them using a few parameters. The research undertaken at Environmental Systems Research Institute (Esri) has focused on developing an increasing set of filters to catch Circular arcs, L-, Spoon-, Door-, U-, Sine-, Z-, Stairs-, Straddle-, and Straight-shapes, etc. All other polyline features not caught by any of these stereotypes fall into the Unknown-shapes. The figures in Figure 2 illustrate the stereotypes that are presently modeled.

The analytical model of the stereotype for an L-shape, for example, is composed of two nearly straight sections connected by a sharp turning point. For Circular arcs, all segments should have their lengths close to an average length and all consecutive turning angles have a same sign and a near average magnitude. A Z-shape stereotype features two consecutive near-90° turning angles in opposite signs, and all polyline sections are near straight. If consecutive near-90° turning angles with alternating signs reach a count larger than 4, a Stairs-shape is formed.

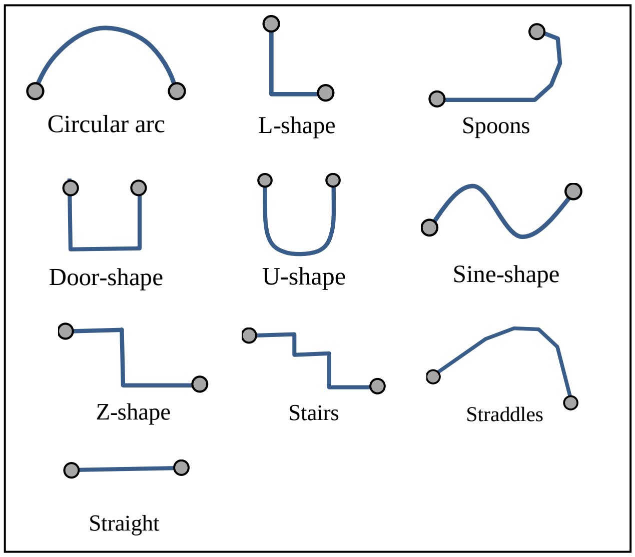


Figure 2. Stereotypes filtering atomic patterns.

Note that real world data of seemingly stereotypic shapes may not all demonstrate expected regularity, stereotype filters need to be able to detect and handle insignificant turbulences in a series of vertices. Techniques such as polyline generalization and statistical deviations can be employed to screen and remove abnormal vertices. Maintaining a robust set of low level pattern filters to hand real data is key to the pattern based feature match method, as the filters will be repeatedly employed in various stages of processing, including to detect non-atomic shapes.

The atomic patterns illustrated in Figure 2 are common shapes that can be observed in most linear geospatial datasets representing urban structures. The set of atomic patterns reflects the maturity of the system in catching geometric shapes using analytic models. New and more complicated patterns would be added as the system evolves.

B. *Composite Patterns*

Composite patterns can be formed by a number of connected polylines, when together they present some distinct figures. In addition to the shapes shown in Figure 2, Circles, Carriageways, Cul-de-sacs, and Straight can be assembled from atomic patterns, as shown in Figure 3.

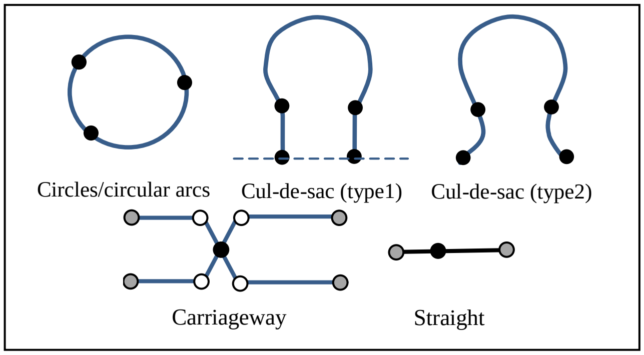


Figure 3. Composite patterns.

A composite Circular arc, for example, can be traced out by looking for consecutive atomic Circular arcs that have a similar curvature and radius. A composite Circle is formed when a Circular arc is closed, else the tracing stops when a currently probed shape does not have similar parameters. A Cul-de-sac can be formed by tracing from both extremes of a Circular arc for a pair of near parallel lines (type1) or a pair of curvature reversed arcs (type2). A Carriageway pattern, generally, involves 4 polylines and is characterized by cross sections joined at a common intersection from which two pairs of near parallel polyline sections are split. Presently, only the above 5 composite structures are assembled in a dataset without referencing structures in a counterpart dataset. Other composite structures will be further formed by reference during the matching process, to be discussed in matching structures.

Forming a composite pattern involves searches at the extending extreme nodes of the current path. It also requires decisions whether a testing atomic shape could be accepted and added into the path. Node topology, path continuity, and pattern compatibility are the factors in the decisions.

C. *Pattern Graph*

A pattern-path-node graph structure is created to collect and operate on the identified patterns, their associative paths, and their geometries, as shown in Figure 4.

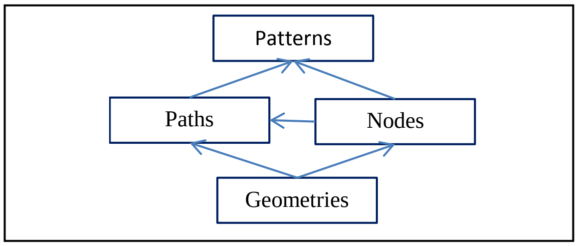


Figure 4. Pattern-path-node graph.

The Geometries are polylines imported from ArcGIS® feature classes, identified by Object ID. Excessive vertices, either duplicated or with distances smaller than a resolution tolerance, are removed through a simplification process. Multipart polylines are consecutively connected as a single shape. Associated with each of the Geometries is a structure holding computed properties to describe the characteristics of a single polyline shape, including the ID of an atomic pattern type. The Nodes are identified as extreme vertices of geometry objects. Associated with each node are IDs of incident Geometries. A path in Paths contains a series of polylines identified by feature IDs that are connected to form a composite structure. A similar data structure to atomic patterns is used to hold characteristic properties of a path. The Patterns is a collection of Paths and associated Nodes that together describe a distinctive structure within a dataset. A pattern object contains a list of main paths that are central to the structure and lists of subordinate paths and nodes that are related to the main paths.

IV. Matching structures

Prior to matching, an inventory of discovered patterns in each graph is established. Furthermore, an indexing of linear features from both graphs is created to facilitate proximity searches. The inventory is used for matching patterns which are listed by pattern types and spatially registered in a coarse grid. The indexing structure is detailed to line segments for refined searches after searches in inventory are firstly attempted. It is convenient to combine segments from both graphs into one indexing structure to pick best matches where multiple candidates are available.

An order of importance will be determined for structure matching, which dictates which pattern type is to be specifically matched first. Our experience reveals that it should proceed from the most complicated to relatively simple structures. This is because simple structures may be part of a larger one. Straight lines will be matched last, just before matching Unknown-shape structures.

For clarity, the first and second datasets involved in matching are named source and target, respectively. The matching process takes a pattern type from source and searches for a counterpart in the target inventory with the closest characteristic values for each of its members. If such a counterpart is found, a match is made. Otherwise, the process will perform proximity searches to identify piecewise shapes from target to fit the larger shape referenced in source. A successful fit will add the composed large structure and modify hierarchical relationships in the target graph. After all known structures in source are exhausted, the process should be repeated for a list of unmatched known shapes from target to match a counterpart from source. It is necessary to repeat matches initiated from either graph, as a large atomic structure existing in one graph may not exist in the other. The reverse process ensures that large known patterns be processed first prior to matching unknown shapes. At the end, structures of both graphs will either find a match or be declared not matched.

Matched pairs may not be exactly of same patterns, but they must be compatible. For example, L- and Spoon-shape patterns are compatible, so are Door- and U-patterns. A straddle may be matched with one of the following compatible shape combinations: a straddle; two spoons; one spoon and one straight; one circular and two straight shapes.

The diagram in Figure 5 illustrates the matching of an atomic straddle shape from the red graph to a number of piecewise polylines in the black graph. The first attempt of finding an atomic straddle counterpart from the black graph is failed. The characteristic sections of the red straddle will be identified, which are a circular arc in the middle part and two near straight sections. Proximity searches then start from the circular section. If a similar circular arc or a spoon shape is returned from the black graph, the rest of the parts will be traced from it. In the example, the search will first find feature 17, which has the closest curvature parameters to the middle section of the red feature 3301. Straight features 12 and 18 are then traced out from 17. Pieced together, the composite straddle in black has properties most similar to that of 3301 in red. A match is done with the red atomic and the black composite straddles.

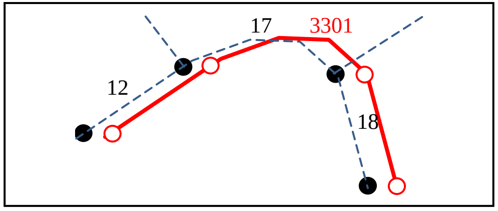


Figure 5. Matching a straddle.

While composing structures during a match, gaps can sometimes exist in one of the graphs, as illustrated in Figure 6. In the example, black circular arcs {1587, 1606, 1608, 1592} form a composite circular path prior to matching. When taking the path to match a counterpart in red graph, two circular paths, {1833} and {1857, 1840} are found, which have similar circular parameters to that of the black. The two red paths together will form a matched composite path, with a gap in between.

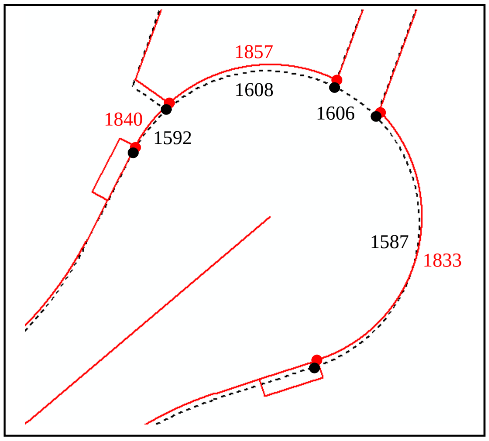


Figure 6. Matching a circular path with gap.

There is also a need to split a previously generated path. This is especially true for large straight paths. Figure 7 illustrates such a case. In the diagram, the black graph shows a straight path composed of shapes {4, 5, 6} and two other straight paths {12} and {9}. The red graph has one red L-shape {64}, and two straight paths {33} and {48}. During the matching process, the red L-shape is matched with {12} and possibly the longer straight path in black. It is necessary to break the long straight into two short ones {4} and {5, 6}.

After breaking, the L-shape is matched with {4, 12} and {5, 6} is added into the black graph structure for further match.

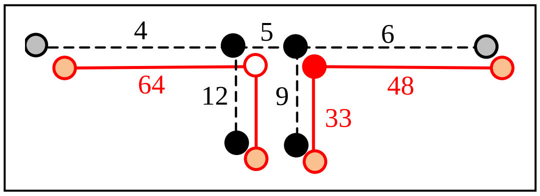


Figure 7. Ilustration of splitting a path.

To match unknown shapes, more comprehensive proximity computations, like buffering, will be needed, since little is known about their characteristics. For each unknown shape, a search for a set of candidates can be constrained by a given or derived search distance, whichever is smaller. The derived search distance can be computed from the result of uniquely matching known structures.

V. Matching Features

A final step with the matched structures is to break them down to feature by feature match. Due to changes in real world and differences in data capturing, features from both graphs will not all have one-to-one correspondences. The following cardinality relationships exist in linear features, as shown in Figure 8. The m:n relationship occurs when it would be ambiguous to break the set of features further down to simpler correspondences. The 1:0 and 0:1 relationships are included to indicate no corresponding features could be matched to satisfy similarity measures.

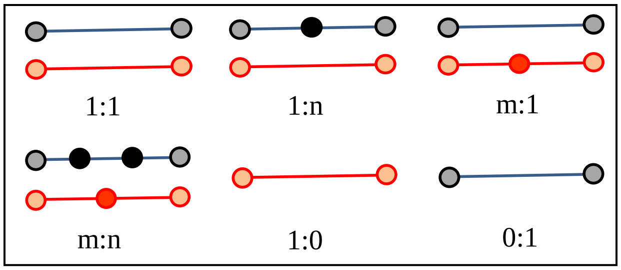


Figure 8. Cardinalities between matched features.

In addition to the length ratio parameter associated with paths, two constraints are considered while matching features from matched structures. The first is topological when other paths incident to nodes and separated by adjacent angles will be analyzed and compared. If there is insufficient topological information, the measure of "orthogonally projected overlapping" between elements of two paths will be applied.

Figure 9 illustrates the process that respects topological measures. Two paths of, source (red) {6, 7, 8} and target (black) {4, 5, 6, 7}, are matched. Features from source will be taken, one at a time, to match features in target. In the diagram, red feature identified as 6 is first matched with black 4. Two determinations will be applied to append black 5 to the match list after 4. First, the length ratio of black 4 is still smaller than that of red 6; second, the extreme node of black 4 is a pseudo node. Adding black 5 to 4 satisfies the length ratio better, furthermore, its front end fits better topologically to the front end of red 6. Other features are matched by applying similar reasoning process using local neighborhood, and considering shape characteristics of incident paths, if necessary.

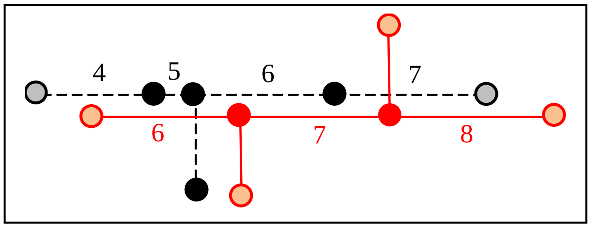


Figure 9. Example of feature matching process.

An orthogonally projected overlapping length is obtained by projecting the extremes of polyline A onto polyline B. If a projection is footed on the extension from an extreme of B, the extreme vertex will be projected back onto polyline A. The projected overlapping length is then calculated with the two points enclosing the orthogonally overlapped section. Figure 10 shows five cases that overlapping lengths are enclosed.

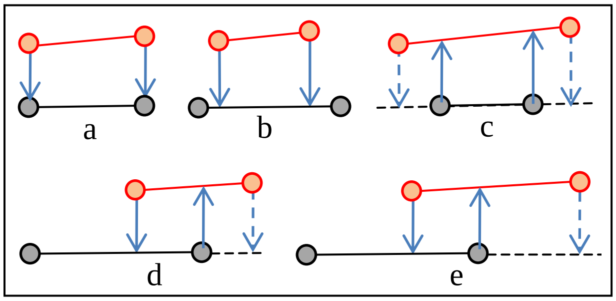


Figure 10. An example of obtaining orthogonally projected length.

The method described in this paper requires that for a pair of features to be considered a match, the projected overlapping length must not be less than half length of the shorter polyline. Case e in Figure 10, for example, does not satisfy the requirement. Using this measure, gaps in the path could correspond to a feature with no match. An example of this case is shown in Figure 6 where the black feature 1606 does not have a sufficient projected overlap with either red 1833 or 1857. It will have a 1:0 match.

VI. Application and Results

The method presented in this paper has been designed as a core component to support three geoprocessing tools for ArcGIS, the commercial GIS software produced and marketed by Esri Inc. They are Detect Feature Changes, Generate Rubbersheet Links, and Transfer Attributes, all of which rely on matching features of two datasets from separate sources covering the same geographic areas.

One of the challenges in developing feature matching techniques, and application tools based on them, is to find effective ways to assess results produced by these tools by feeding user data of various complexities. The evaluation is necessary for users to gain confidence on the levels of accuracy that could be expected from using these tools. Another challenge is to find practical workflows to complete conflation tasks with high levels of automation.

Applying the above mentioned conflation tools to real world data, Lee, Yang, and Ahmed [12] have devised workflows to perform data integration tasks such as spatial adjustment, transferring attributes, and generating reports on feature changes by detecting spatial and attribute discrepancies. Furthermore, a set of script tools, written with Python or built by chaining other tools in ArcGIS, have been developed to automatically verify and evaluate outputs produced by the conflation tools. Their assessment shows an achievement of above 90% of feature matching and conflation accuracy in executing the workflows on top of multiple user datasets demonstrating excellent, ordinary, and poor similarities. The successful rates are compatible to those claimed in published papers [4], [6]. Due to unavailability of software developed based on the other published feature matching methods, a cross-comparison under similar conditions on performance, ease of use, and robustness, etc., cannot be reported in this paper.

VII. Conclusion and Future Work

A method of matching linear features overlapping the same geographic area has been described. By catching atomic and composite feature patterns to construct a pattern graph, better understanding of a dataset is obtained. Patterns of a dataset are recognized through a set of stereotypes as low level constructs, which can be applied to compose large pattern structures within a dataset and with reference to the other dataset during the matching process. Based on the graphs built on source and target data, matching features starts with matching structures, in which locating of paired structures becomes less dependent on coordinates, rotation, and shift, but more on referencing local neighborhood structures. Processes of matching individual features are explained. Consideration factors, determining how matched structures are broken down into matched features, are also elaborated. The method is implemented as a core component which is used for producing conflation geoprocessing tools in ArcGIS. Testing and application of the tools in practical workflows have demonstrated promising results.

While the research reported in this paper has established a framework in developing feature matching based tools, more work is needed to complete the missing parts of the methodology. First, a full analysis on the algorithms in accomplishing matching features is necessary in terms of time and space complexity, from which comparisons to other methods could be made. Evidence of practical uses, and results from applying the conflation tools and workflows in solving real world problems, should be included as an integral part of the method. Meanwhile, it is anticipated that the method and its applications will continue to evolve on the following fronts:

- Maintaining existing pattern recognition stereotypes so that they become more versatile and robust;
- Developing new patterns to reduce the number of unknown-shape elements in graphs;
- Developing reasoning on top of identified geometric patterns to enrich semantic meanings, hence the metadata of a dataset;
- Considering matching patterns coming from datasets with varying generalization scales; and
- Researching on geospatial matching and conflation between vector datasets and raster images.

Acknowledgment

Many members in the Esri Geoprocessing team have contributed to the conflation project through discussions.

Some diagrams contain sections of LA parcel data from the Department of Public Works, Los Angeles County, USA.

References

[1] M.P. Lynch and A. Saalfeld, “Conflation: Automated Map Compilation - A Video Game Approach,” AUTOCARTO 7 Proceedings, 1985, pp. 343- 352.

[2] A. Samal, S. Seth, and K. Cueto, “A feature-based approach to conflation of geospatial sources,” International Journal of Geographical Information Science, 18:5, 2004, pp. 459-489.

[3] A. Schwering, “Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey,” Transactions in GIS, 2008, 12(1), pp. 5–29.

[4] M. Zhang and L. Meng, “Delimited stroke oriented algorithm-working principle and implementation for the matching of road network,” Geographic Information Sciences, Hong Kong, June 2008, 14(1), pp. 44-53.

[5] V. Walter and D. Fritsch, “Matching spatial data sets: a statistical approach,” International Journal of Geographical Information Science, 13:5, 1999, pp. 445-473.

[6] L. Li and M. F. Goodchild, “An optimisation model for linear feature matching in geographical data conflation,” International Journal of Image and Data Fusion, vol. 2:4, 2001, pp. 309-328.

[7] G. Touya, A. Coupé, J. L. Jollec, O. Dorie, and F. Fuchs, “Conflation Optimized by Least Squares to Maintain Geographic Shapes,” ISPRS Int. J. Geo-Inf. 2013, 2, pp. 621-644, doi:10.3390/ijgi2030621.

[8] W. Mackaness and G. Edwards, “The importance of modelling pattern and structure in automated map generalisation,” Joint ISPRS/ICA Workshop: Multi-Scale Representation of Spatial Data. Ottawa, Canada, July 7-8, 2002, pp. 1-12.

[9] S. Marshall, “Streets and patterns,” 1st ed., London and New York: Spon Press, 2005, ISBN 0-415-31750-9.

[10] F. Heinzle, K.-H. Anders, and M. Sester, “Automatic detection of patterns in road networks – methods and evaluations,” 2007, Proceedings of Joint Workshop Visualization and Exploration of Geospatial Data, Stuttgart, vol. XXXVI-4/W45 (CD-ROM).

[11] B. Jiang and X. Liu, “Automatic generation of the axial lines of urban environments to capture what we perceive,” International Journal of Geographical Information Science, 2010, 24(4), pp. 545–558.

[12] D. Lee, W. Yang, and N. Ahmed, “Conflation in geoprocessing framework – case studies,” submitted to GEOProcessing, 2014, Barcelona, Spain.

The Compression Algorithm of Hyperspectral Space Images Using Pre-byte Processing and Intra-bands Correlation

Alexander Zamyatin
Institute of Cybernetics of National Research
Tomsk Polytechnic University
Tomsk City, Russia
e-mail: zamyatin@tpu.ru

Assiya Sarinova
Innovation Eurasia University
Pavlodar City, Kazakhstan
e-mail: assiya_prog@mail.ru

Pedro Cabral
Institute of Statistics and Information Management
New University of Lisbon
Lisbon, Portugal
e-mail: pcabral@isegi.unl.pt

***Abstract* — This paper concerns the compression of hyperspectral data from Earth remote sensing by suggesting a multistage algorithm to increase the compression ratio, using a formation of auxiliary data of the large redundancy in their byte presentation, and taking in account the correlation intra-bands. Here are presented the results of the studies on the effectiveness of compression of hyperspectral space images made by the proposed compression algorithm with the universal and specialized compression software.**

***Keywords* - remote sensing; hyperspectral images; lossless compression; intra-bands correlation; byte representation of data.**

I. Introduction

Modern centers for space monitoring and systems of Earth Remote Sensing (RS) continually process, archive and distribute data which constitute tens and hundreds of gigabytes [1-9]. A key problem in the process is compressing RS data to increase effectiveness of a data transfer via connection channels of limited carrying capacity and archiving in RS storage subsystems of a limited capacity. The necessary classification of this data must be of the highest value. That is why it is more appropriate lossless compression, which is free of any distortions of statistic brightness characteristics of restored data.

The solution of the compression problem, which is the most accessible for practical implementation presupposes the usage of universal and widely-known algorithms and means of compression, for instance, in the archival software *WinRar*, *WinZip* or compressor *Lossless JPEG (JPEG-LS)* on the base of *JPEG* [10-15] image compression standard. However, RS data is classified by various characteristics – spectral, radiometric, spatial resolutions and by geometrical size of the scene. The above-mentioned universal means of compression fail to consistently consider the variable differences [16-19]. Thus, there are multispectral and hyperspectral Aerospace Images (AI), which have essentially different parameters of spectral resolution and are characterized by dependence (correlation) between data of different bands [20]. If for multispectral aerospace images the coefficient of mutual correlation of bands is R∈(0.3; 0.8), then for hyperspectral AI representing values of brightness which are received in various spectral bands with the high spectral resolution, the correlation of neighboring bands is R ≈ 1.0. This demonstrates high redundancy of data and expediency of applying this feature while compression [21][22].

Besides, knowing the correlation value (intra-bands correlation) between bands of hyperspectral AI, it is expedient to operate values of deviations (difference) between it and actual reference values, and that will allow to reduce the range of data change and hence will demand a smaller number of categories for their storage. The effective usage of specialized means of the compression is possible, considering the above-stated features of hyperspectral AI.

The software intended for the analysis of spatial data and processing of AI are *ERDAS Imagine*, *ERDASER Mapper*, *ArcView GIS*, *GeoExpress* and others often dispose of specific modules for compression of AI. Thus, the package *ERDAS Imagine* utilizes the compression tool for images in the format of MrSID (*IMAGINE MrSID Desktop Encoder* and *IMAGINE MrSID Workstation Encoder*), based on wavelets and intended for lossy compression of large RS images. The system *ERDASER Mapper* has modules of loss compression on the basis of the standards *JPEG2000* and *ECW* [23]. In the package *ArcView GIS* there is a module *MrSID* which compresses raster images files with loss [24]. The software product *GeoExpress* is intended for compressing raster data with the use of popular formats *MrSID* and *JPEG2000* [25]. Thus, all commercial systems of RS data processing are have all means of loss compression on the basis of well-known standards.

Today hyperspectral RS images lossless compression research attempt to apply various approaches and methods [26-34]. The separate stages of transforming data while compressing AI and possibility of decreasing power inputs and algorithmic complexity are discussed. Additionally, there are various attempts at adapting standards which have successfully proved for compressing usual images for compressing hyperspectral AI.

ISBN: 978-1-61208-326-1

II. DESCRIPTION OF THE ALGORITHM

However, not all details of the original algorithms of compression are clear. Their numbers exceed the possibilities of most widespread means of compression; when applied to hyperspectral AI with various characteristics is uncertain.

It is our purpose to promote further search for approaches to lossless compression of hyperspectral RS images, substantially free from the disadvantages of existing universal and specialized means of compression.

Considering the features of hyperspectral AI and some details of existing analogues, the most expedient solution of the problem of compressing hyperspectral AI is by multi-stage transformations: first the advantages of universal traditional approaches for data compression, and secondly to consider the specificity of hyperspectral data. The algorithm embodying this approach and some of its results is discussed below.

Considering the specificity of compressing hyperspectral AI, the proposed algorithm has the following stages:

1. To consider the functional dependence of values of brightness (albedo) between various bands of images, by calculating the correlation and of deviations (differences) of initial data and the values of those found for functional dependence.
2. Creation of auxiliary structure of data on the basis of the initial hyperspectral AI, storing the unique pair groups of values of elements in a byte representation, and addressing references on these unique pair groups as well.
3. Compression of received data transformations with standard entropy algorithm by processing the generated auxiliary structures of the data.

Let us consider the details of the above-mentioned stages.

In the first stage, the value of deviations of the linear dependence on the matrix of values **I**[*m*,*n*,*k*].

For step-by-step description of the first stage consisting in searching correlation and deviations it is necessary to account the following objects:

- An initial image **–** the matrix of values of the image **I**[*m*,*n*,*k*], where *m*,*n*,*k* – are indices of the lines, columns and bands of the initial image, *m* = 1,2,…,*M*, *n* = 1,2,…,*N*, *k* = 1,2,…,*K*;
- ***R***[*k*] – a file for "level-by-level" preservation values of correlation *R* between the neighboring bands (layers);
- **Q**[*k*] – a file for placing values of the mathematical expectation for each band **I**[*m*,*n*,*k*];
- **L**[*k*] – a file for preserving the linear dependence;
- **I'**[*m*,*n*,*k*] c a file for placing values of differences (deviations) between **L**[*k*] and **I**[*m*,*n*,*k*].

Step 1. To calculate a mathematical expectation m_k of each band of the initial image **I**[*m*,*n*,*k*] and to place values to the file **Q**[*k*], as in (1):

$$\mathbf{Q}[k]=\sum_{k=1}^{K}\mathbf{I}[m,n,k]\times\rho_{\mathbf{I}}{}^{k} \qquad (1)$$

where $\rho_{\mathbf{I}}{}^{k}$– relative frequency of occurrence of values of the image **I**, *k*=1,2,…, *K*.

Step 2. To calculate (on the base of **Q**[*k*] and **I**[*m*,*n*,*k*]) correlation *R* for each pair of all available *K* bands of the initial image **I**[*m*,*n*,*k*], as in (2):

$$\mathbf{R}[k]=\frac{\sum_{\mathrm{m,n}}(\mathbf{I}[m,n,k]-\mathbf{Q}[k])}{\sqrt{\sum_{\mathrm{m,n}}(\mathbf{I}[m,n,k]-\mathbf{Q}[k])^{2}}} \qquad (2)$$

To place the result to **R**[*k*], *k*=1,2,…, *K*-1.

Step 3. To calculate (on the base of **Q**[*k*] and **R**[*k*]) a linear dependence of the kind $L=m_k\times R$ for each pair of the available *K* bands of the initial image **I**[*m*,*n*,*k*]. Result should be placed to **L**[*k*], **L**[*k*] = **Q**[*k*]×**R**[*k*], *k*=1,2,…,*K*-1.

Step 4. To calculate (on the base of **R**[*k*] and **L**[*k*]) the difference between elements **L**[*k*] and corresponding values of the initial image **I**[*m*,*n*,*k*] in each of *K* bands, as in (3):

$$\mathbf{I}'[m,n,k]=\begin{cases} by\ \mathbf{R}[k]>0,\mathbf{I}[m,n,k]-\mathbf{L}[k-1] \\ by\ \mathbf{R}[k]\le 0,\mathbf{I}[m,n,k] \end{cases} \qquad (3)$$

for *m*=1..*M*, *n*=1..*N*. The result is placed to **I'**[*m*,*n*,*k*].

Step 5. To transform the negative values **I'**[*m*,*n*,*k*] to positive ones in the byte representation as numbers with a sign demand more bytes than those without a sign (4):

$$\mathbf{I}'[m,n,k]=\begin{cases} \text{at } \mathbf{I}[m,n,k]\ge 0,\ 2\times\mathbf{I}[m,n,k] \\ \text{at } \mathbf{I}[m,n,k]=0,-2\times\mathbf{I}[m,n,k]-1 \end{cases} \qquad (4)$$

The result is the file **I'**[*m*,*n*,*k*] considering values of intra-bands correlation ***R***[*k*].

The essence of the second stage consists of forming a file of unique pair groups of values which represent the initial image in the byte representation. Then, the file is formed containing references to the same pair group of values.

The algorithm proceeds with two additional objects:

- **M**[*j*,*k*] – a file of unique pair groups of values of the initial image in the byte representation;
- **D**[*j*,*k*] – a file for entering references (to unique pair groups of values).

In second stage of transformation, the step-by-step elaboration is shown in Figure 1:

Step 1. To form (on the base of **I'**[*m*,*n*,*k*]) a file **M**[*j*,*k*] for each band *K* adding unique pair groups of values in the byte representation from the file **I'**[*m*,*n*,*k*]. To place the result to **M**[*j*,*k*], *j*= 1,2,..,*J*. If repeated pair groups of values are absent, then *J = (M*×*N*×*K*)/2, *k*= 1,2,..,*K*.

Step 2. To form (on the base of **M**[*j*,*k*]) a file **D** [*j*,*k*], putting down references to unique pair groups of values from the file **M**[*j*,*k*] to **D**[*j*,*k*]. To place the result to **D**[*j*,*k*], *j*<*J* as *j* = *M*×*N*×*K*, *k*= 1,2,..,*K*.

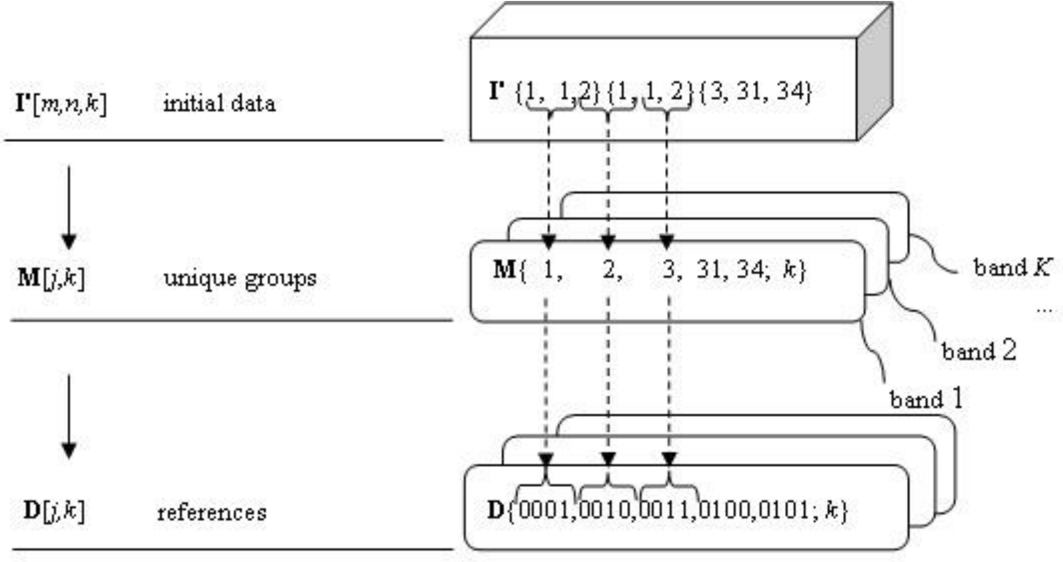


Figure 1. Procedure of forming auxiliary structure of data.

At the end of the second stage, context modeling was used with the known arithmetic coding for compressing data of the file **D**[*j*,*k*] to the archival software **D'**[*j*,*k*].

In order to form an initial hyperspectral AI **I**[*m*,*n*,*k*] from **D'**[*j*,*k*] it is necessary to make a number of transformations opposed to the above-mentioned:

- To make arithmetic decoding of the file **D'**[*j*,*k*] restoring the file **D**[*j*,*k*];
- To find in the file **D**[*j*,*k*] the corresponding references to unique pair groups from the formed structure of the data **M**[*j*,*k*];
- To restore the file **I'**[*m*,*n*,k] containing a file of dependencies **L**[*k*] having counted the absolute values of the file **I'**[*m*,*n*,k] and having restored the initial image **I**[*m*,*n*,k].

III. EXPERIMENTAL RESEARCH

In order to access the effectiveness of the proposed algorithm in what concerns both the point compression ratio as the limits of its application, a number of experiments was perfomed using hyperspectral AI of the system RS *AVIRIS* (table 1) in the format of data of raster geoinformation system *Idrisi Kilimanjaro*. The system *AVIRIS* (*Airborne Visible/Infrared Imaging Spectrometer*) provides 224 spectral images with the wavelength of the band from 400 nanometers to 2500 nanometers. Also, the comparison of the proposed algorithm with the results of experiments received for universal archivers compression algorithms *WinRar*, *WinZip* and compressor *Lossles JPEG* which applies the resolution of compression standard *JPEG* widely used in commercial compression systems were made.

The experiments are made on a computer with the processor *Intel Core i5* 2,5 GigaHerz and RAM 4 Gigabit under operating system *Windows 7* (*updating package 3*).

TABLE I. EXAMPLES OF CHARAC TERISTICS OF TEST DATA (OF THE SYSTEMS RS *AVIRIS*)

Name of hyperspectral AI	K	M × N, pixel	File size, byte
f970619t01p02_r07_sc01	224	100 × 100	6140096
f970619t01p02_r07_sc02	224	200 × 200	36199296
f970619t01p02_r07_sc03	224	300 × 300	81178496
f970619t01p02_r07_sc04	224	400 × 400	144077196
f970619t01p02_r07_sc05	224	500 × 500	210746396
f970619t01p02_r07_sc06	224	624 × 512	281673728

The proposed algorithm was then performed sequence of stages to find the most effective:

Sequence I – Consideration of the correlation (1) → forming an auxiliary of data with unique pair groups and references to them (2) → arithmetic coding (3).

Sequence II – Forming an auxiliary of data with unique pair groups and references to them (1) → arithmetic coding (2).

Sequence III – Consideration the correlation (1) → arithmetic coding (2).

Sequence IV – Forming an auxiliary of data with unique pair groups and references to them (1) →consideration the correlation (2) → arithmetic coding (3).

In order to evaluate the most productive sequence considering the contribution of each of these stages a number of experiments with different variants was conducted (Figure 2).

ISBN: 978-1-61208-326-1

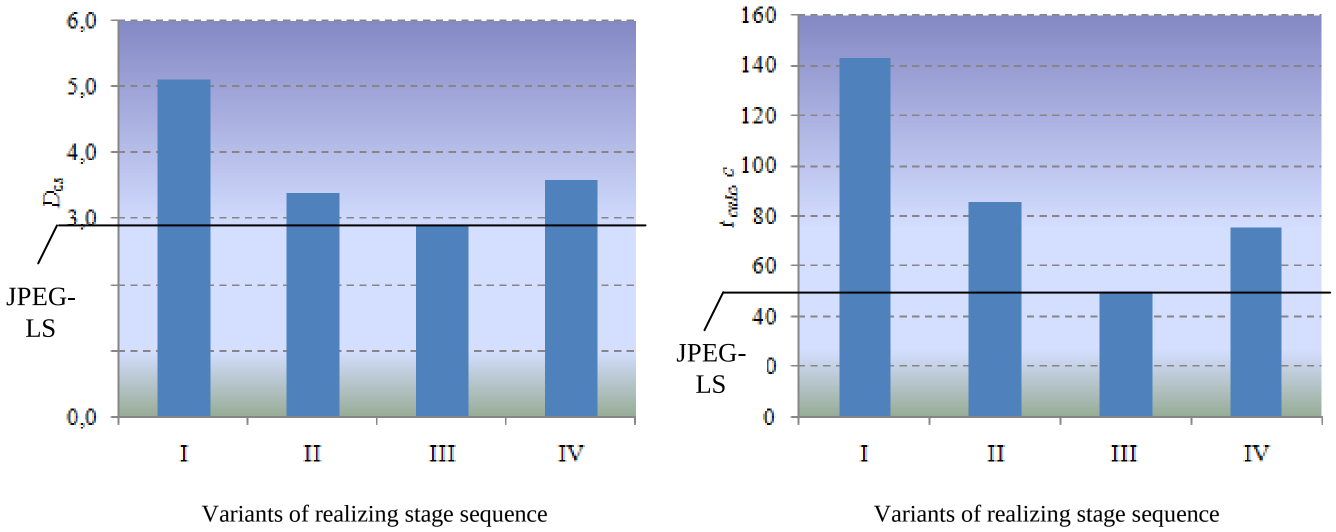


Figure 2. Comparison of indicators of variants of realizing a compression algorithm: a) by exponent of compression D_{cs}; b) by time of calculating work t_{calc}

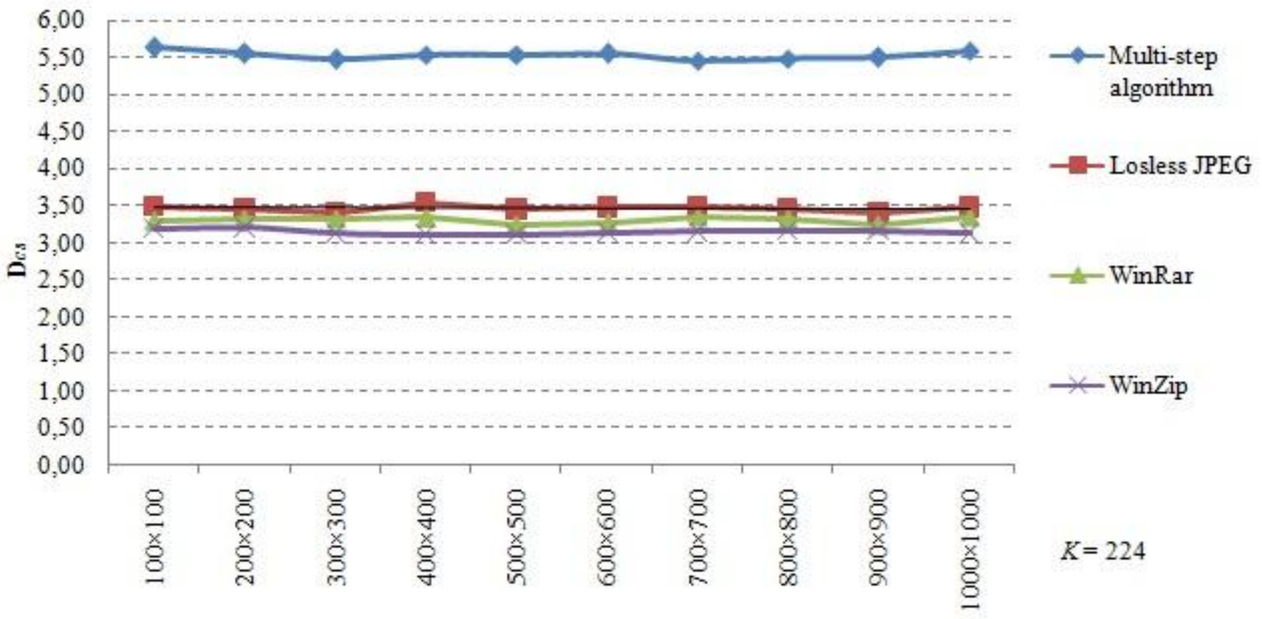


Figure 3. Comparative effectiveness of compression algorithms for different geometrical size of scene.

A fragment of the _results of this experiment is displayed results in the Figure 2, that shows that various stages of the algorithm have different importance while forming a result. In the 1st, 2nd and 4th variants of the stage sequences, the results surpass *Losless JPEG* indifferent degrees (from 25% to 46%).

The best result is achived by sequence I is with the highest exponent of compression. Sequence IV was impaired by the execution of stage it is not expedient to realize the stage (of considering the correlation after the stage of forming auxiliary structures of data) as it leads to increasing a range of data change while counting the difference. The results of realizing the second stage show that absence of the stage of intra-bands correlation leads to an insignificant exponent of compression in comparison with *Losless JPEG*.

Realization of the stage sequence in accordance with the sequence III leads to the most insignificant result as there is no powerful formation of unique pair groups and references to them with creation of corresponding auxiliary structures of data.

In Figure 3, the results of comparative experiments demonstrate the superiority of the proposed algorithm over analogues in exponent of compression D_{cs} at the varied geometrical sizes of hyperspectral AI. At increasing the geometrical size of the scenes, all the investigated algorithms show a steady result which exhibits little to no change.

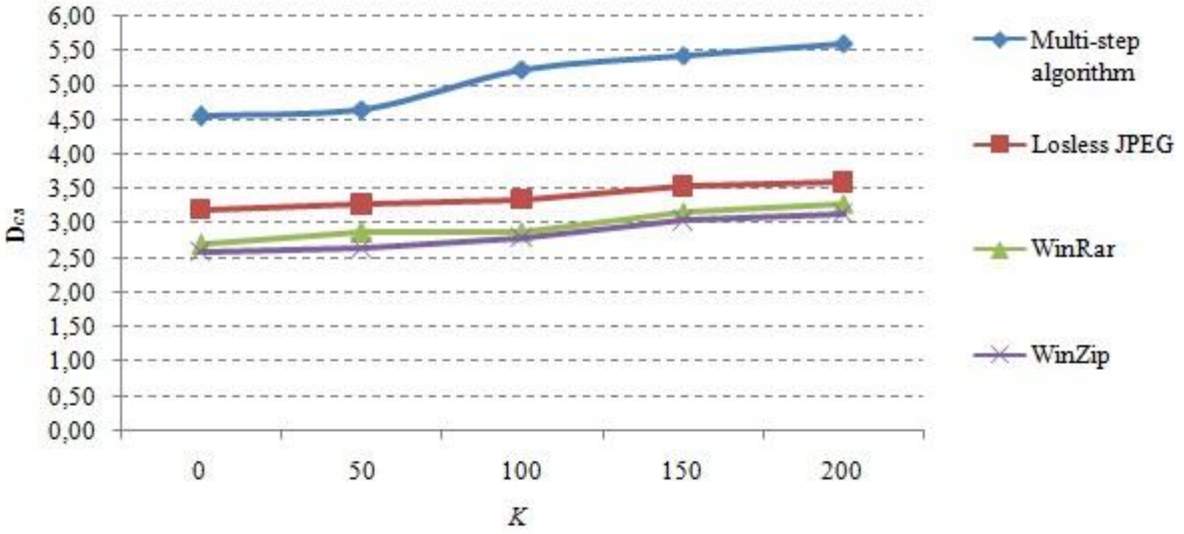


Figure 4. Dependence of compression exponent D_{cs} of algorithms on the number of bands *K*.

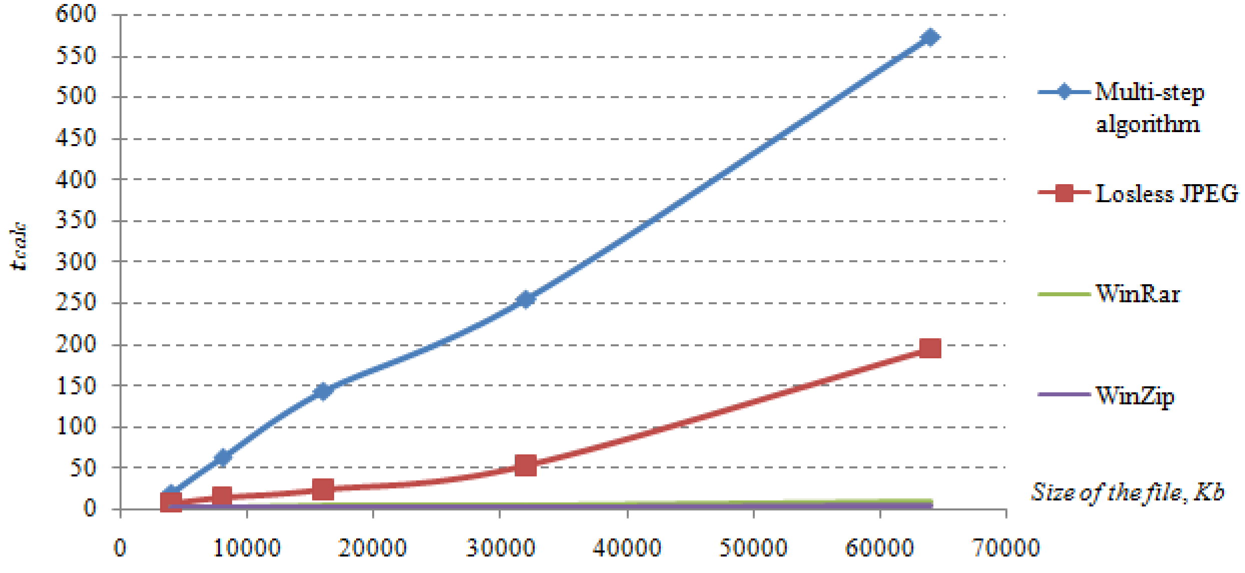


Figure 5. Computational performance of compression algorithms.

Research was conducted to explore the dependence of compression exponent D_{cs} on the number of bands of AI *K* (Figure 4). Results show that compression exponent D_{cs} is increasing proportionally to the number of bands *K*as the redundancy of the data of hyperspectral AI raises.

In conducting comparative research of compression exponents it is necessary to pay attention to calculating expenses of compression algorithms (Figure 5).

As seen in Figure 5, in the proposed algorithm, the calculated effectiveness in comparison with analogues increased 3 fold. This is explained by an improved multi-stage algorithm which is provided to form auxiliary structures of data on AI considering the correlation and the following arithmetic coding. Universal archival software does not take into account the specificity of the data being compressed and does not account for such operations.

IV. CONCLUSION

1. A multi-stage algorithm for compressing hyperspectral AI was developed. This algorithm considers intra-bands correlation and the preliminary byte data processing, allowing up to 46 % increase in data compression when compared with other algorithms.

2. The analysis of the importance of stages has shown that the stage of preliminary byte processing with formation of auxiliary structures of data allows to improve the result considerably – up to 45 %. The stage of considering intra-bands correlation is less significant. However it allows to lower a range of varied values for operating by smaller spacing, allowing to increase the compression ratio considerably – up to 26 %.

3. The analysis of computing efficiency has shown, that in order to achieve significant results of compression in applying a multi-stage algorithm, high computing expenses are required, conceding to the nearest analogue *Lossless JPEG* up to 3 folds.

This work has been supported by Russian Foundation for Basic Research (grant *№ 14-07-00027a*).

REFERENCES

[1] V.G. Bondur, “Modern approaches to processing of hyperspectral space images”,Research Institute of space monitoring “Aerospace”, Moscow, 2013, p. 4.

[2] M.A. Popov and S.A. Stankevich, “Optimization methods of spectral bands number in tasks of processing and data analysis of distance remote sensing of the earth”, Scientific center of aerospace earth research, 2003, no. 1, pp. 106-112.

[3] H. Wang, S. D. Babacan, and K. Sayood, “Lossless hyperspectral-image compression using context-based conditional average”, vol. 45, no. 12, 2007, pp. 4187–4193

[4] H. Chengfu, R. Zhang, and P. Tianxiang, “Lossless Compression of Hyperspectral Images Based on Searching Optimal Multibands for Prediction”, vol. 6, no. 2, 2009, pp. 339–343.

[5] Y. Liang, L. Jianping, and Ke G., “Lossless compression of hyperspectral images using hybrid context prediction”, vol.20, no. 7, 2012,. pp. 199–206.

[6] B. Aiazzi, L. Alparone, S. Baronti, C. Lastri, and M. Selva, “Spectral Distortion in Lossy Compression of Hyperspectral Data”, Journal of Electrical and Computer Engineering, no. 10, 2012, p. 8., doi:10. 1155/ 2012/ 850637

[7] L. Cheng-chen and H. Yin-tsung, “Lossless Compression of Hyperspectral Images Using Adaptive Prediction and Backward Search Schemes”, Journal of Information Science and Engineering, no. 27, 2011, pp. 419–435.

[8] B. Aiazzi, L. Alparone, and S. Baronti, “Near-lossless image compression by relaxation-labeled prediction, Signal Process”, no. 11, 2002, pp. 1619–1631.

[9] E. Magli, G. Olmo, and E. Quacchio, “Optimized onboard lossless and near-lossless compression of hyperspectral data using CALIC”, IEEE Geoscience and remote sensing letters, no. 1, 2004, pp. 21–25.

[10] B. Aiazzi, S. Baronti, and L. Alparone, “Lossless compression of hyperspectral images using multiband lookup tables”, IEEE Signal Process. Letters, vol.6, no. 16, 2009, pp. 481–484.

[11] B. Penna, T. Tillo, E. Magli, and G. Olmo, “Transform Coding Techniques for Lossy Hyperspectral Data Compression”, IEEE Geoscience and remote sensing letters, vol. 45, no. 5, 2007, pp. 1408–1420.

[12] RarLab. WinRar software system for compress files [http://www.win-rar.com/rarproducts.html]. Retrieved: 15.01.2014.

[13] WinZip. Program of compression for Windows [http://www.winzip.com/ru/prodpagewz.htm]. Retrieved:. 19.01.2014

[14] X. Tang, W. Pearlman, and J. Modestino, “Hyperspectral image compression using three-dimensional wavelet coding”, Proc. SPIE IS&T, 2003, pp. 1037–1047.

[15] B. Penna, T. Tillo, E. Magli, and. G. Olmo, “Progressive 3-D coding of hyperspectral images based on JPEG 2000”, IEEE Geoscience and remote sensing letters. vol. 1, no. 3, 2006, pp. 125–129.

[16] J. Zhang and G. Liu, “An efficient reordering prediction-based lossless compression algorithm for hyperspectral images”, IEEE Geoscience and remote sensing letters.–. vol. 2, no. 4, 2007, pp. 283–287.

[17] J. S. Mielikainen, A. Kaarna, and P. Toivanen “Lossless hyperspectral image compression via linear prediction”, Proc. SPIE 4725, no. 8, 2002, pp. 600–608.

[18] F. Rizzo, B. Carpentieri, G. Motta, and J. A. Storer, “Low-complexity lossless compression of hyperspectral imagery via linear prediction”, IEEE Signal Process. Lett, vol. 2, no. 12, 2005, pp. 138–141.

[19] ISO/IEC 15444-1. JPEG2000 Image Coding System [http:/www.jpeg.org/public/15444-1annexi.pdf]. Retrieved: 10.01.2014.

[20] A. Kiely, M. Klimesh, H. Xie, and N. Aranki, “ICER–3D: A Progressive Wavelet-Based Compressor for Hyperspectral Images”, 2006, pp. 142–164.

[21] L. Gueguen, M. Trocan, B. Pesquet-Popescu, A. Giros, and M.A. Datcu, “Comparison of multispectral satellite sequence compression approaches”, Signals, Circuits and Systems, no. 1, 2005, pp. 87–90.

[22] A.V. Zamyatin and T. D. Chung, “Compression of multispectral space images using wavelet transform and intra-bands correlation”, Journal of Tomsk Polytechnic University, vol. 313, no. 5., 2008, pp. 20–24.

[23] Interregional public organization promoting the market development of geographic information technologies and services “GIS-Association”, [http://www.gisa.ru/1489.html], Retrieved: 15.01.2014.

[24] Arc View GIS [http://gisa.ru/3577.html]. Retrieved: 10.01.2014.

[25] Geo-information systems [http://loi.sscc.ru/gis/default.aspx]. Retrieved: 05.01.14.

[26] D. Vatolin, A. Ratushnyak, M. Smirnov, and V. Yukin, ”Data compression methods”, M.:Dialogue, MIFI, 2003, p. 384.

[27] V.N. Kopylov, “Creation basics of aerospace environment monitoring”. Yekaterinburg: PP “Kontur”, 2006, p. 144.

[28] V.F. Babkin, I.M. Knizhny, and K.E. Chrekin, “Compression of multispectral images with out loss or limited losses’, Modern and perspective developments and technologies in a space instrumentation: Moskva, IKI RAN,. no. 1, 2004, pp. 330–332.

[29] F. Rizzo, B. Carpentieri, G. Motta, and J.A. Storer, “Low-complexity lossless compression of hyperspectral imagery via linear prediction”, IEEE Signal Process. Letters, vol. 2, no. 12, 2005, pp. 138–141.

[30] G. Motta, F. Rizzo, and J.A. Storer, “Hyperspectral Data Compression”, Berlin: Springer, 2006, p. 415.

[31] H. Goto, Y. Hasegawa, and M. Tanaka, “Efficient Scheduling Focusing on the Duality of MPL Representatives,” Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, 2007, pp. 57–64, doi:10.1109/SCIS.2007.357670.

[32] D. Salomon and G. Motta, Hand book of data compression, 5th ed., Springer London Dordrecht Heidelberg New York, 2010, p. 42, doi:10. 1007/10.1007/978-1-84882-903-9

[33] A. Robert, “Remote Sensing Models and Methods for Image processing”, Department of Electrical and computer Engeneering College of optical Sciences, University of Orizona, USA, 2007, p. 592.

[34] L. Chang and C. M. Cheng, and Y.L. Chang, “Fast Hyperspectral Image Classification Using Binary Quaternion-Moment-Preserving Thresholding Technique”, IEEE International Geoscience and Remote Sensing Symposium, 2008, Boston, U.S.A., p.2.

Knowledge Processing for Geosciences, Volcanology, and Spatial Sciences Employing Universal Classification

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU),
Leibniz Universität Hannover,
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

***Abstract*—This paper presents the results from creating and using long-term knowledge resources for knowledge processing by employing a universal classification. The knowledge objects can further comprise specialised research data collections and refer to any kind of multi-disciplinary data. The resources can be sustainably used for documentation, universal classification, and structuring as well as with scientific supercomputing resources for advanced information systems, supporting discovery and decision making. The case study discusses selected examples from the geosciences context. Within the case studies the structure of the resources and the Universal Decimal Classification (UDC) has been used together with spatial information in order to handle and support workflows and visualisations for multi-disciplinary data and features. Implementations of various components have been created for discovery and development using dynamical, interactive, and batch computing and storage resources in an Integrated Information and Computing System environment exploiting High End Computing (HEC) and High Performance Computing (HPC) resources and UDC. The paper discusses the new results and practical cases regarding a multi-disciplinary geosciences, volcanology, and spatial features scenario from long-term knowledge resources.**

Keywords*–Geosciences; Knowledge Resources; Integrated Systems; Sustainability; Information Systems; Volcanology; Archaeology; Classification; UDC; High Performance Computing.

I. INTRODUCTION

Geoscientific knowledge processing is traditionally focussed on processing and analysis of data resulting from geophysical or geological measurements. Examples are processing based on seismological, seismic, magnetic, or gravimetric data. The amount of information and documentation from geosciences and natural sciences based methods and features as well as their complexity has steadily increased for decades. Efficiency and economical practice forces to long-term document and exploit this pool of multi-disciplinary information. Spatial and chronological data and classification are an indispensable component. It is becoming increasingly important that with most professional analysis different geophysical methods and results have to be used in combination.

Common means of application and knowledge discovery, e.g., isolated batch or interactive application scenarios or string based search routines on plain data cannot even approximately integrate the required higher complexity of real environments.

The knowledge gathered during generations should be considered the most valuable component, the more important for long-term results from geosciences. The universal knowledge resources require long-term documentation as well as universal classification and structuring, beyond traditional collections, digital libraries and isolated content [1], [2]. With the long-term multi-disciplinary resources the high end processing and computing aspects are essential for sustainability and discovery. Therefore, it is recommended to implement scientific supercomputing resources supporting advanced information systems and creating and improving workflows as recommended [3] with Integrated Information and Computing System (IICS) components and High End Computing (HEC) [4]. This paper presents the results from creating and managing long-term knowledge resources for knowledge processing by employing a universal classification like the Universal Decimal Classification (UDC) [5]. It discusses the experiences handling systematics and classification as well as the methodological use of “Object Carousels”. The paper points out the demands and challenges as resulting from the case studies within the GEXI collaborations [6] concentrating on integrating knowledge from geosciences, volcanology, and spatial sciences disciplines.

This paper is organised as follows. Section II introduces the previous work and components used, Sections III and IV present the systematics and classification used for the processing, and discuss the results from the implementation case study, Sections V and VI evaluate main results, and summarise the lessons learned, conclusions and future work.

II. COMPONENTS EMPLOYED

The data used here is based on the content and context from the LX Foundation Scientific Resources [7], [8]. The LX structure and UDC [5], [9] are an essential means for the processing workflows and evaluation of the knowledge objects and containers. The applied workflows and processing are based on the data and extended features developed for the Gottfried Wilhelm Leibniz resources [10]. The classification is state-of-the-art within the knowledge resources, which implicitly means that the classification is not created statically or even fixed. It can be used and dynamically be modified on the fly, e.g., when required by a knowledge discovery workflow description. Representations and references can be handled dynamically with the context of a discovery process. So, the classification can be dynamically modelled with the context of the workflow. The LX resources can provide any knowledge documentation and additional information on objects as well as, e.g., geo- and knowledge references. The volcanological data used in

ISBN: 978-1-61208-326-1

the examples is embedded into millions of multi-disciplinary objects, dynamical and spatial information and data files.

The knowledge objects are under continuous development for more than twenty-five years. The classification information has been added in order to describe the objects with the ongoing research and in order to enable more detailed documentation in a multi-disciplinary context. The knowledge resources can make sustainable and vital use of Object Carousels [11] in order to create knowledge object references and to modularise the required algorithms [12]. This provides a universal means for improving coverage, e.g., dark data, and quality within the workflow. Therefore, for the cases presented in this research paper we had to concentrate on the structure of the knowledge objects, georeferencing, and references data. As secondary components, besides IICS applications and interfaces are available, allowing parallel workflows and intelligent components on HEC and HPC resources [13], [14]. With the IICS, the Generic Mapping Tools [15] have been used to visualise georeferenced data wherever a spatial representation is reasonable.

III. Disciplines, systematics, and processing

In geosciences, there is no globally unique stratigraphy. Different continents and regions require different and detailed stratigraphies. Therefore, it is not practicable to have a flat unique global standard due to the regional differences in geological development. Present common stratigraphy concepts [16] fail on general use as well as on a consistent universal classification required. Implementing a universal long-term use we further need to consider appropriate systematics, e.g., lithostratigraphical, chronological, biostratigraphical, chronometrical, chronostratigraphical systematics. For example when it comes to plants, animals, and genotype "-zoic", "-phytic" or "-gen" often mix without distinction. Instead of a mix-up of terminology, for systematical use the alignment of the Eonothems/Eons, Erathems/Eras, Systems/Periods, Series/Epochs, and Stages/Ages and so on should be handled consistently and consequently. In addition, the multi-regional dimension should be available for these, showing correspondence with the appropriate absolute ages, as available on-site.

For an efficient and effective processing the knowledge data requires a flexible structure and a universal systematic classification. Any knowledge resources documenting complex multi-disciplinary reality for discovery applications require features for exact documentation on the one hand and they require soft criteria on the other hand. UDC is a classification complying with the classification criteria. Together with the content, which may deliver more detail or differing perspectives UDC provides a universal view on the classified objects. When requiring facetted classification for multi-disciplinary knowledge the universal UDC cannot be ignored as it is the most comprehensive and flexible means available and supported. With the knowledge resources in this research handling 70000 classes, for 100000 objects and several million referenced data the algorithms are mostly non-linear. They allow interactive use, dynamical communication, computing, decision support, and pre- and postprocessing, e.g., visualisation.

The classification deployed for documentation [17] is able to document any object with any relation, structure, and level of detail as well as intelligently selected nearby hits and references. Objects include any media, textual documents, illustrations, photos, maps, videos, sound recordings, as well as realia, physical objects such as museum objects. UDC is a suitable background classification, for example: The objects use preliminary classifications for multi-disciplinary content. Standardised operations used with UDC are coordination and addition ("+"), consecutive extension ("/"), relation (":"), order-fixing ("::"), subgrouping ("[]"), non-UDC notation ("*"), alphabetic extension ("A-Z"), besides place, time, nationality, language, form, and characteristics.

IV. Case study implementation and results

The following sections discusses the work done for using knowledge resource objects with processing and computing from within IICS. For knowledge resources it is necessary that any classification can be added while the content is developed, over long period of time, more than decades. With the cases presented the content has been created over more than twenty-five years. Methodologically, in a first phase objects have been documented without classification. In a second phase, all objects describing volcanic features have been classified as volcanic features. In a third phase, volcanic features' objects have been classified into separate classes as required with ongoing extended description of objects in a multi-disciplinary context. The case study presents a state-of-the-art selection of volcanological and geological features. An evaluation of the association that users have, showed that the criteria "date" and "location" are most prominent with objects if the workflow approaches from the "surface (of the earth)" view [6]. Mapping and timelining will be the natural result.

1) Space: Table I shows an excerpt of the resulting UDC classification practically used for spatial features and place implemented with knowledge resources and geo-coordinates.

Table I. Universal Decimal Classification of spatial features and place used with the knowledge resources (excerpt).

UDC Code	*Description*
UDC:(1)	Place and space in general. Localization. Orientation
UDC:(1-0/-9)	Special auxiliary subdivision for boundaries and spatial ...
UDC:(1-0)	Zones
UDC:(1-1)	Orientation. Points of the compass. Relative position
UDC:(1-19)	Relative location, direction and orientation
UDC:(1-2)	Lowest administrative units. Localities
UDC:(1-5)	Dependent or semi-dependent territories
UDC:(1-6)	States or groupings of states from various points of view
UDC:(1-7)	Places and areas according to privacy, publicness ...
UDC:(1-8)	Location. Source. Transit. Destination
UDC:(1-9)	Regionalization according to specialized points of view
UDC:(2)	Physiographic designation
UDC:(20)	Ecosphere
UDC:(21)	Surface of the Earth in general. Land areas in particular. ...
UDC:(23)	Above sea level. Surface relief. Above ground generally. ...
UDC:(24)	Below sea level. Underground. Subterranean
UDC:(25)	Natural flat ground (at, above or below sea level). ...
UDC:(26)	Oceans, seas and interconnections
UDC:(28)	Inland waters
UDC:(29)	The world according to physiographic features
UDC:(3)	Places of the ancient and mediaeval world
UDC:(32)	Ancient Egypt
UDC:(36)	Regions of the so-called barbarians
UDC:(37)	Italia. Ancient Rome and Italy
UDC:(38)	Ancient Greece
UDC:(4/9)	Countries and places of the modern world

2) Time: Table II shows an excerpt of the resulting UDC classification of spaces of time practically used with the knowledge resources. Instead of the earlier UDC editions the classifications are composite UDC:551.7 mappings referring to historical geology and stratigraphy for all the spaces of time.

TABLE II. UNIVERSAL DECIMAL CLASSIFICATION OF SPACES OF TIME USED WITH THE KNOWLEDGE RESOURCES (EXCERPT).

UDC Code	*Description*
UDC:”0/2”	Dates and ranges of time (CE or AD) ...
UDC:”0”	First millennium CE
UDC:”1”	Second millennium CE
UDC:”2”	Third millennium CE
UDC:”3/7”	Time divisions other than dates in Christian ...
UDC:”3”	Conventional time divisions and subdivisions ...
UDC:”4”	Duration. Time-span. Period. Term. Ages ...
UDC:”5”	Periodicity. Frequency. Recurrence at ...
UDC:”6”	Geological, archaeological and cultural time divisions
UDC:”61/62”	Geological time division
UDC:”63”	Archaeological, prehistoric, protohistoric periods ...
UDC:”67/69”	Time reckonings: universal, secular, non-Christian ...
UDC:”67”	Universal time reckoning. Before Present
UDC:”68”	Secular time reckonings other than universal and ...
UDC:”69”	Dates and time units in non-Christian ...
UDC:”7”	Phenomena in time. Phenomenology of time
UDC:551.7+”61”	Cryptozoic aeon. Precambrian. 600+ MYBP ...
UDC:551.7+”616”	Archaean. Ur-gneiss formation. Ur-schiefer formation
UDC:551.7+”618”	Eozoic. Algonkian
UDC:551.7+”62”	Phanerozoic aeon. 600 MYBP - Present
UDC:551.7+”621”	Palaeozoic. 600-180 MYBP
UDC:551.7+”621.2”	Cambrian. 600-490 MYBP
UDC:551.7+”621.3”	Ordovician. 490-430 MYBP
UDC:551.7+”621.4”	Silurian. Gothlandian. 430-400 MYBP
UDC:551.7+”621.5”	Devonian. 400-350 MYBP
UDC:551.7+”621.6”	Carboniferous. 350-270 MYBP
UDC:551.7+”621.7”	Permian. 270-220 MYBP
UDC:551.7+”622.2”	Triassic. 220-180 MYBP
UDC:551.7+”622.4”	Jurassic. 180-135 MYBP
UDC:551.7+”622.6”	Cretaceous. 135-70 MYBP
UDC:551.7+”628”	Cenozoic (Cainozoic). Neozoic
UDC:551.7+”628”	Tertiary. 70-1 MYBP
UDC:551.7+”628.2”	Palaeogenic. Nummulitic
UDC:551.7+”628.22”	Palaeocene
UDC:551.7+”628.24”	Eocene
UDC:551.7+”628.26”	Oligocene
UDC:551.7+”628.4”	Neogene
UDC:551.7+”628.42”	Miocene
UDC:551.7+”628.44”	Pliocene
UDC:551.7+”628.6”	Quaternary. 1 MYBP - Present
UDC:551.7+”628.62”	Pleistocene in general. Diluvium
UDC:551.7+”628.64”	Holocene. Postglacial in general

Any of the classification can be mapped to specific content data. The workflows and processing handle different dates and specification between classification and content as well as using equal classification elements for different absolute dates, e.g., as required for different regions or cultures.

3) Results of systematical use: Suitable views for volcanic features are: Type (of volcano, coarse categories), date on timeline, size (height). For craters respective views are: Type (of crater, fragmentary), date on timeline, size (diameter). Two Object Carousels have been computed. Figure 1 shows the knowledge resources groups for volcano types, and Figure 2 provides the geological spaces of time references. For simplicity only the main groups are shown, subgroups like for Quarternary “Holocene” and “Pleistocene” create separate carousels (Figure 3). Most geological objects have references into some instance of these carousels. This enables to create numberless links to additional information.

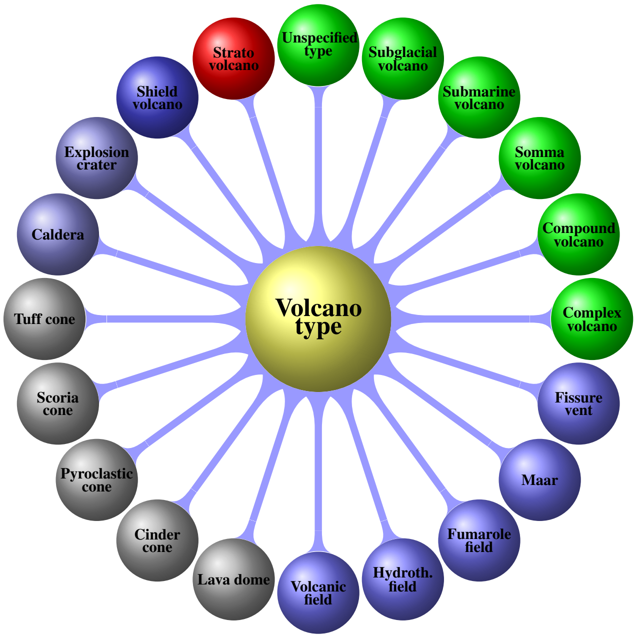


Figure 1. Object Carousel for volcano and type references computed for terrestrial volcanism, providing volcano type references.

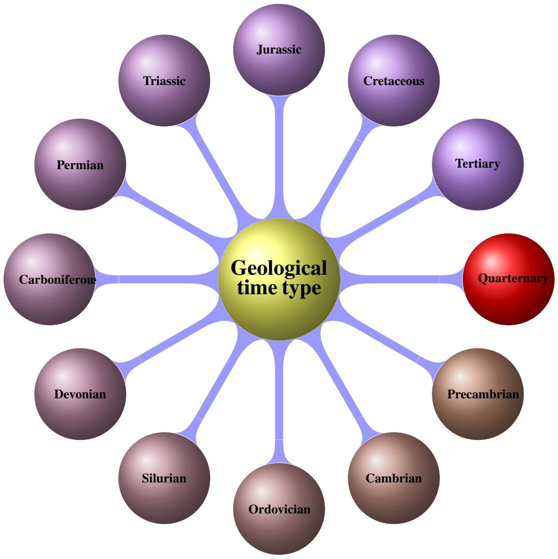


Figure 2. Object Carousel on geological spaces of time for computed references (terrestrial volcanoes, impact craters, and geological processes).

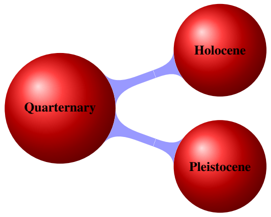


Figure 3. Object Carousel “Quarternary”.

The colour coding for Carousels is symbolic and can be defined to represent any grouping as decided within the workflow. It can result from the grade of detail required for the description. In this case, the colour red links the three shown

Object Carousels with the information referring to a requested object like “Vesuvius”. The subgroup Object Carousels, e.g., “Quarternary” (Figure 3), opens additional references to volcanological feature objects. The listing in Figure 4 shows context replacement definitions and corrections.

```
Cretacious :: Cretaceous
Kreide :: Cretaceous
Trias :: Triassic
Carbon :: Carboniferous
Karbon :: Carboniferous
Silurium :: Silurian
Silur :: Silurian
Ordovicium :: Ordovician
Ordovizium :: Ordovician
Cambrium :: Cambrian
Kambrium :: Cambrian
Precambrium :: Precambrian
Präkambrium :: Precambrian
```

Figure 4. Replacement definition for relevant terms (LX resources).

The example lists an excerpt of relevant terms and types of notation that can be considered equal for the target context.

4) Processing media citation references: Figure 5 shows an excerpt of a media citation set used with UDC classified knowledge objects, here with a Vesuvius reference.

```
cite: YES 20070000 {LXK:Pompeii; Vesuvius; reconstruction
 ; 3D; animation; Holocene} {UDC:...} {PAGE:----..----}
 LXCITE://Bonaventura:2007:My_DVD
cite: YES 20130000 {LXK:Pompeii; Vesuvius; Vesuvio;
 Holocene; postcard} {UDC:...} {PAGE:----..----} LXCITE:
 //Guardasole:2013:Vesuvio_1270m
cite: YES 20070000 {LXK:Pompeii; Vesuvius; reconstruction
 ; diorama} {UDC:...} {PAGE:----..----} LXCITE://
 Bonaventura:2007:Pompeii
cite: YES 20070000 {LXK:Pompeii; Vesuvius; bakery; mill
 stones; material; stone; volcanic lava; basalt; Holocene
 ; diorama} {UDC:...} {PAGE:--56..--59} LXCITE://
 Bonaventura:2007:Pompeii
```

Figure 5. Media citation set excerpt used with the UDC classified knowledge object “Vesuvius” (LX resources).

The examples are part of the “Vesuvius” and “volcanic mill stone” object references. The media citations refer to 3D video animations and dioramic reconstructions as well as even to postcards. These references resolve to [18], [19], [20].

5) Classification development: All classifications are subject of a continuous development, review, and auditing process. Table III shows an example in different UDC editions.

TABLE III. DEVELOPMENT OF “TERTIARY” CLASSIFICATION WITH UDC EDITIONS AND KNOWLEDGE RESOURCES (EXCERPT).

UDC Code (a)	*UDC Code (b)*	*Description*
UDC:”623”	UDC:”628”	Tertiary (70-1 MYBP)
UDC:”623.1”	UDC:”628.2”	Palaeogene (70-25 MYBP)
UDC:”623.5”	UDC:”628.4”	Neocene (25-1 MYBP)
UDC:551.77	UDC:551.7+”628”	Cenozoic (Cainozoic). Neozoic
UDC:551.78	UDC:551.7+”628”	Tertiary. 70-1 MYBP
UDC:551.781	UDC:551.7+”628.2”	Palaeogenic. Nummulitic
UDC:551.781.3	UDC:551.7+”628.22”	Palaeocene
UDC:551.781.4	UDC:551.7+”628.24”	Eocene
UDC:551.781.5	UDC:551.7+”628.26”	Oligocene
UDC:551.782	UDC:551.7+”628.4”	Neogene
UDC:551.782.1	UDC:551.7+”628.42”	Miocene
UDC:551.782.2	UDC:551.7+”628.44”	Pliocene

The example is the “Tertiary” classification development within different UDC editions. The table shows that the target not only moved $(a) \rightarrow (b)$ within the classification but was also adapted to a new subgrouping (lower block). The currently final result is a composite classification, composing from geology and time, holding both Tertiary and Cenozoic.

UDC still not considers different stratigraphies in plain. Figure 6 shows Object Carousels computed for a complete common system (top) as well as for an alternative system (below) used for some purposes [16] after the year 2000, missing “Tertiary”. The colours represent the term levels within the respective system.

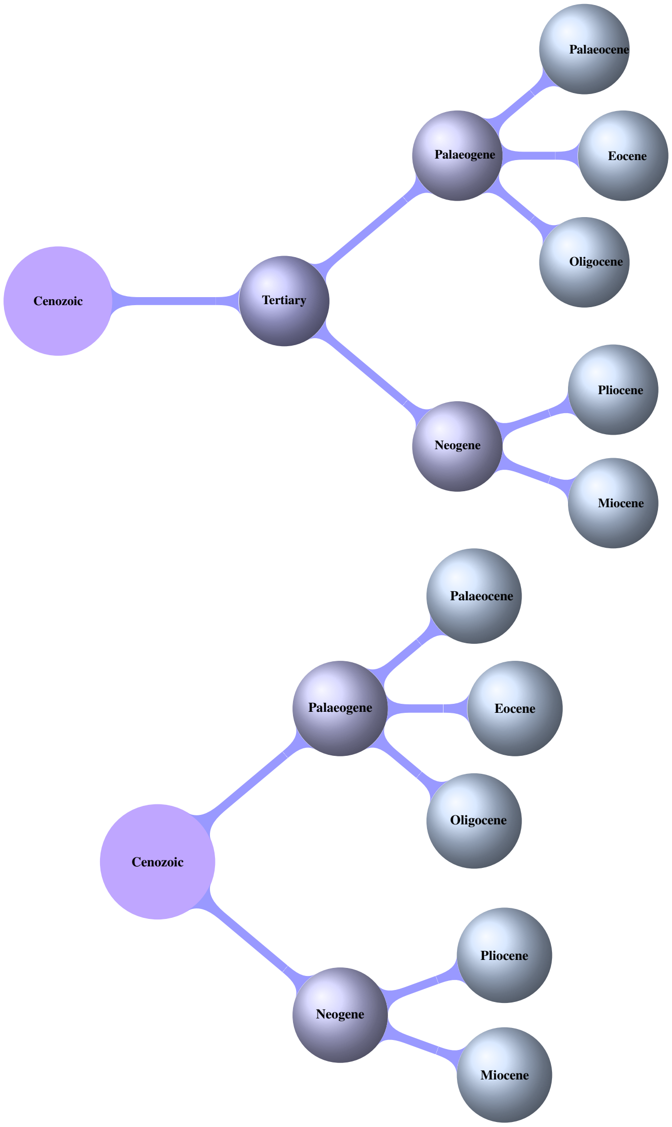


Figure 6. Object Carousel “Tertiary”: Common (top) and alternative (below).

Moved items have to be considered “persistent” within long-term knowledge resources appropriately with all consequences. It is possible to support any number of versions within the knowledge resources as long as each is handled consistently.

6) Result matrix: Table IV shows the results from the computation of a systematical classification of volcanological features, short “volcano types”.

TABLE IV. COMPUTED SYSTEMATICAL CLASSIFICATION OF VOLCANOLOGICAL FEATURES FROM THE KNOWLEDGE RESOURCES.

Volcano Type	*Group*	*References Data Examples*
Complex volcano	A	Vesuvius VNUM:0101-02= UDC:[551.21+911.2+55]:[902]”63”(4+23+24)... GPS:40.821N14.426E Quarternary VEI:VEI5
Compound volcano	A	Cayambe VNUM:1502-004 UDC:[551.21+911.2+55]:(8+23+24)... GPS:... Holocene ...
Somma volcano	A	Ebeko VNUM:0900-38= UDC:... GPS:... Quarternary ...
Submarine volcano	A	Campi Flegrei Mar Sicilia VNUM:0101-07= UDC:... GPS:... Quarternary ...
Subglacial volcano	A	Katla VNUM:1702-03= UDC:... GPS:... Quarternary ...
Unspecified type	A	– VNUM:– GPS:... – ...
Strato volcano	B	Vulcano VNUM:0101-05= UDC:... GPS:... Quarternary ...
Shield volcano	C	Etna VNUM:0101-06= UDC:... GPS:... Quarternary ...
Explosion crater	D	Larderello VNUM:0101-001 UDC:... GPS:... Quarternary ...
Caldera	D	Campi Flegrei VNUM:0101-01= UDC:... GPS:... Quarternary ...
Tuff cone	E	Tutuila VNUM:0404-02- UDC:... GPS:... Holocene ...
Scoria cone	E	Antofagasta de la Sierra VNUM:1505-124 UDC:... GPS:... Holocene ...
Pyroclastic cone	E	Anunciacion, Cerro VNUM:1405-032 UDC:... GPS:... Holocene ...
Cinder cone	E	Chiquimula Field VNUM:1402-20- UDC:... GPS:... Holocene ...
Lava dome	E	El Chichon VNUM:1401-12 UDC:... GPS:... Quarternary ...
Volcanic field	F	Holotepec VNUM:1401-07- UDC:... GPS:... Quarternary ...
Hydrothermal field	F	Musa River VNUM:0503-02= UDC:... GPS:... Quarternary ...
Fumarole field	F	Kos VNUM:0102-06= UDC:... GPS:... Pleistocene ...
Maar	F	West Eifel Volcanic Field VNUM:0100-01- UDC:... GPS:... Quarternary ...
Fissure vent	F	Quetena VNUM:1505-074 UDC:... GPS:... Holocene ...

It compiles a small excerpt of computed data from the LX resources [7]. The table delivers comprehensive information for the volcanological topics integrated here: Volcanic feature types, computed groups, UDC mappings, and examples of computed references, e.g., Volcano Number (VNUM) the volcanic reference file number, geo-coordinates and spatial data, and spaces of time, as well as referenced data, e.g., the Volcanic Explosivity Index (VEI) [21]. The full result matrix for this request contains several hundreds of computed objects with tenthousands of references. A container represents a collection of equally structured groups of related objects on a certain topic. In addition, in depth completion within object containers has been enabled for the case of volcanological features. The resources further allow for a flexible mapping of attributes, e.g., container relations, classification, keywords, numbers, references, media samples, material samples, spatial data, and geological spaces of time. With these references the volcanological features can be associated with a VEI, e.g., Vesuvius (Pompeii) VEI5, Krakatau VEI6, Tambora VEI7, Thera (Santorini) VEI7, Toba (Sumatra) VEI8, whereas a "Caldera" object itself being a crater does not have a VEI. With existing models used in simulation and modelling there is no consideration of references between disciplines, e.g., volcanoes and weather. With the knowledge resources, volcanological features can be referred to volcanological events, seismological events, and weather phenomenon events or biology. The larger the data base is the more correlatable events get available in space and time. In comparison to mono-disciplinary information the multi-disciplinary context of the knowledge resources supports an improved knowledge description. Further, even indirect correlation, e.g., in the above case between volcanic features and meteorite impact features can be investigated.

7) Knowledge generation, combination, and visualisation: The following visualisation (Figure 7) paradigmatically illustrates the results from the compute requests. An on-location attribute has been choosen for the relations in order to compute a distribution map for volcanic features using the `lxlocation` workflow. The location attribute is suitable for referring to an unlimited number of multi-disciplinary information in this case.

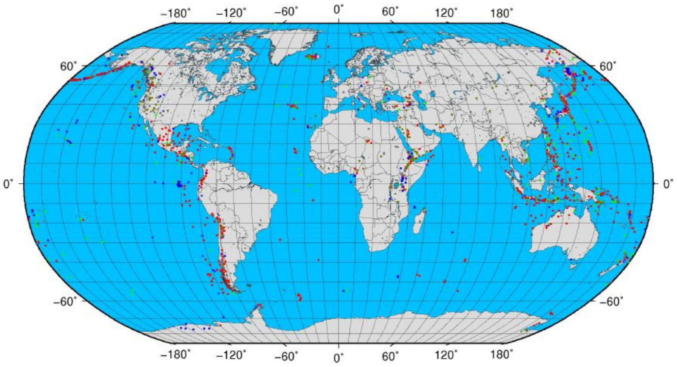


Figure 7. Volcanomap – computed worldwide spatial distribution of classified volcanological features from resulting object entries.

The distribution is computed from the result matrix of related object context of several hundred classified terrestrial volcanic features via the knowledge resources research database. The result matrix is the result of the present content, references, and workflow. In all examples only an excerpt of these can be shown. Several modules have been used for this example: `select_knowledge_environment`, `lxgrep_udc`, `lxkwgrep`, `generateCarousel`, `lxvolcanoes2gmt`, `cprgmt_world_cprvolcanoes_separated`, as well as `pscoast`, `pstoraster`, and `psxy`. System interfaces can be created via instructions, programming interfaces, or any kind of interface the disciplines working on implementations and suggested workflows want to built on top of the knowledge resources. The workflow allows any feature supported by the deployed components, e.g.,

- Association by classification weighting,
- Association by grouping,
- Association by colourisation,
- Association and by symbolisation.

In this example, colour groups have been computed via the result matrix (Table IV): A: green, B: red, C: blue, D: lighter blue, E: grey, F: dark green. The volcanic features are classified and several classification groups have been choosen for the result. The map shows the present situation according to the present state of the available volcanological data. It is possible to combine any information, e.g., computing a map animation varying in time, showing the development of volcanic features.

8) Processing and computational numbers and issues: Table V shows the processing and computational demands per instance resulting from the presented scenarios.

TABLE V. PROCESSING AND COMPUTATIONAL DEMANDS.

Item	*Value / Description*
UDC, number of classification items	70,000
Number of classification languages	50
Number of classification variations (50×70000)	3,500,000
Knowledge object subset, number of items	100,000
Number of terms	10,000,000
Number of object languages	2
Operations, number per subset result entry	50,000,000
Number per subset result entry, incl. keywords	500,000,000
Parallelisation (subset), wall time / num. of nodes	7,500 s / 1
Wall time / number of nodes	1,300 s / 10
Wall time / number of nodes	220 s / 100
Wall time / number of nodes (extrapolated)	4 s / 10,000

Besides the large requirements per instance with most workflows there are significant effects by parallelising even within single instances. The following issues have shown to lead to advanced challenges and increased processing and computational times. Nomenclature, terms, and attributes tend to be at least partially different in different cultures and languages. For many discovery workflows as well as efforts to increase the quality of the result matrices it is necessary to consider more than one culture and language. Processing a classification numbering in decreasing numbering with increasing age or following in different directions is less consequent. For example, in geosciences it is natural to start spaces of time with Quarternary, followed by older stratigraphy. In addition to the existing singular spaces of time mapping most objects require appropriate different mappings to absolute dates, e.g., with Bronze Age having different absolute dates for different regions or cultures. The calculation with extensive composite classifications, facets, and respective ranges instead of native classifications can increase the computational requirements drastically as has been shown with the knowledge content from the Gottfried Wilhelm Leibniz resources [10].

V. DISCUSSION AND EVALUATION

The Knowledge Oriented Architecture (KOA) of the resources is based on a flexible integration of the documentation and development architecture utilising the Collaboration house framework for disciplines, services, and resources [8]. The knowledge objects, here the geological and volcanic feature objects, can be used with any of their attributes. Therefore, any references to objects belonging or referring to any other objects can be computed from this. For an object referred to a timescale of periods other objects can be associated with the respective object, even beyond direct references. For example, “geological time type” can refer from “volcano type” to any other suitable for a geological or comparable spaces of time classification. This will, e.g., be true for geophysical, palaeontological or archaeological objects. Further, volcanic objects from the Quarternary can be associated to meteorite impact events from the Quarternary. The more, they can be restricted to associated objects of a certain attribute, e.g., from the same region. With secondary steps further information can be integrated. This can include geophysical data, media data or associated objects. The resulting quality depends on an intelligent use of context and classification. A strong classification support is essential, the more as object and even many citations, media, and publications are not explicitly aware of the nomenclature of spaces of time used with specific content can, e.g., to express that the spaces of time refer to plants or animals. Employing a universal classification with multi-disciplinary content this way, e.g., with volcanological content, expedites knowledge discovery as well as it targets on scientific discovery. Regarding methodology it further allows to

- Support a systematic documentation,
- Define a normative classification,
- Define cognitive interfaces.

Regarding architecture and implementation it allows to

- Support decision making in complex systems,
- Implement learning system components and
- Support components by intelligent systems.

Creating classified knowledge resources objects has proven to be most sustainable for a significant period of operation and implementation. It has been efficient and portable with all application scenarios and environments for more than two decades, used with ten different operating system environments, with different editing components, processing languages, and compute and storage resources. From classification side it is suggested to have advanced computing support, e.g., for spaces of time as well as for the complementary systematics for disciplines. In addition, a methodological framework for UDC supporting the required processing and computation would add immense benefits to its universal applicability. Some new types of stratigraphies have not widely been adopted and should again become subject of modification regarding a long-term use. In many cases, the consequence of claims on consistency has been to use one dedicated edition of the classification. This shall ensure consistency within the application. Using a small subset of classification can help to reduce the appearent work that has to be done for classification but it cannot ensure to avoid variances in different editions. Consistent version management support for the classification has shown to become necessary as soon as knowledge resources are using modified classifications over time.

VI. CONCLUSION AND FUTURE WORK

The knowledge processing employing UDC classification has shown to be a universal and most flexible solution for creating long-term multi-disciplinary knowledge resources. The resources and framework can be used even with basic attributes and cross-references, and assure support for subsequent use and knowledge procurement processes. Structuring and classification with long-term knowledge resources and UDC support have successfully provided excellent solutions, which can be used for natural sciences, e.g., geosciences, volcanology or with spatial disciplines as well as for universal knowledge. The knowledge resources can provide any kind of Object Carousel and object references. Decisions can be computed with support of the UDC classification. Due to the universal long-term multi-disciplinary knowledge gathering, the knowledge resources are a general universal decision support base.

Besides these, a major benefit of the extensive support of UDC language translations is that regarding discovery workflows it can also be used for improving the quality as well as the quantity of elements and references in the result matrices. Employing a universal classification when creating knowledge resources has provided substantial benefit for both. The workflow procedures build for special purposes are property of the researchers and disciplines creating, developing, and operating their implementations. The data used by them is intended to be part of the respective collaboration. Currently, if someone creates data, he or she can use the data and share it with others, creating agreements and policies. As knowledge resources have been proven to be a valuable means for research in many disciplines, components are candidates for community tasks as well as for open access development and licensing models. Currently the policies with many collaborations, funding, and services (as comparable with the UDC model) do not allow to make sources and content public. Because the process of creating long-term sustainable content is quite pretentious and will never be completed there might be support by a sustainable funding in the future, too.

As presented, the knowledge processing can base on a solid and sustainable long-term resource, which allows to create any kind of workflows, dynamical discovery, and IICS components and facilitate the use of High End Computing resources.

ACKNOWLEDGEMENTS

We are grateful to all national and international partners in the GEXI cooperations for the innovative constructive work. We thank the Science and High Performance Supercomputing Centre (SHPSC) for long-term support of collaborative research since 1997, including the GEXI developments and case studies on archaeological and geoscientific information systems. Special thanks go to the scientific colleagues at the Gottfried Wilhelm Leibniz Bibliothek (GWLB), Hannover, and at the Institute for Legal Informatics (IRI), Leibniz Universität Hannover, and WWU Münster, as well as to the participants of the EULISP Programme for prolific discussion of scientific, legal, and technical aspects over the last years.

REFERENCES

[1] "The Digital Archaeological Record (tDAR)," 2014, URL: http://www.tdar.org [accessed: 2014-01-12].

[2] "WDL, World Digital Library," 2014, URL: http://www.wdl.org [accessed: 2014-01-12].

[3] Wissenschaftsrat, "Übergreifende Empfehlungen zu Informationsinfrastrukturen, (English: Spanning Recommendations for Information Infrastructures)," *Wissenschaftsrat, Deutschland, (English: Science Council, Germany), Drs. 10466-11, Berlin, 28.01.2011*, 2011, URL: http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf [accessed: 2014-01-12].

[4] C.-P. Rückemann, *Queueing Aspects of Integrated Information and Computing Systems in Geosciences and Natural Sciences*. InTech, 2011, pp. 1–26, Chapter 1, ISBN-13: 978-953-307-737-6, DOI: 10.5772/29337.

[5] "Universal Decimal Classification Consortium (UDCC)," 2014, URL: http://www.udcc.org [accessed: 2014-01-12].

[6] "Geo Exploration and Information (GEXI)," 1996, 1999, 2010, 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/index.html#GEXI [accessed: 2014-01-12].

[7] "LX-Project," 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/#LX (Information) [accessed: 2014-01-12].

[8] C.-P. Rückemann, "Enabling Dynamical Use of Integrated Systems and Scientific Supercomputing Resources for Archaeological Information Systems," in *Proc. of The Int. Conf. on Advanced Communications and Computation (INFOCOMP 2012), October 21–26, 2012, Venice, Italy*. XPS, 2012, pp. 36–41, ISBN: 978-1-61208-226-4.

[9] "UDC Online," 2014, URL: http://www.udc-hub.com/ [accessed: 2014-01-12].

[10] C.-P. Rückemann, "Archaeological and Geoscientific Objects used with Integrated Systems and Scientific Supercomputing Resources," *International Journal on Advances in Systems and Measurements*, vol. 6, no. 1&2, pp. 200–213, 2013, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), URL: http://www.thinkmind.org/download.php?articleid=sysmea_v6_n12_2013_15 [accessed: 2014-01-12].

[11] C.-P. Rückemann, "Sustainable Knowledge Resources Supporting Scientific Supercomputing for Archaeological and Geoscientific Information Systems," in *Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal*. XPS Press, 2013, pp. 55–60, ISSN: 2308-3484, ISBN: 978-1-61208-310-0.

[12] C.-P. Rückemann, "High End Computing for Diffraction Amplitudes," in *Proceedings ICNAAM 2013, Rhodes, Greece*, vol. 1558. AIP Press, 2013, pp. 305–308, ISBN: 978-0-7354-1184-5, ISSN: 0094-243X, DOI: 10.1063/1.4825483.

[13] U. Inden, D. T. Meridou, M.-E. C. Papadopoulou, A.-C. G. Anadiotis, and C.-P. Rückemann, "Complex Landscapes of Risk in Operations Systems Aspects of Processing and Modelling," in *Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal*. XPS Press, 2013, pp. 99–104, ISSN: 2308-3484, ISBN: 978-1-61208-310-0.

[14] P. Leitão, U. Inden, and C.-P. Rückemann, "Parallelising Multi-agent Systems for High Performance Computing," in *Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal*. XPS Press, 2013, pp. 1–6, ISSN: 2308-3484, ISBN: 978-1-61208-310-0.

[15] "GMT - Generic Mapping Tools," 2014, URL: http://imina.soest.hawaii.edu/gmt [accessed: 2014-01-12].

[16] International Commission on Stratigraphy, "International Chronostratigraphic Chart," 2014, URL: http://www.stratigraphy.org/ICSchart/ChronostratChart2013-01.pdf [accessed: 2014-01-12].

[17] C.-P. Rückemann, "Integrating Information Systems and Scientific Computing," *International Journal on Advances in Systems and Measurements*, vol. 5, no. 3&4, pp. 113–127, 2012, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), URL: http://www.thinkmind.org/index.php?view=article&articleid=sysmea_v5_n34_2012_3/ [accessed: 2014-01-12].

[18] M. My, *Pompeii Reconstructed (DVD)*. ArcheoLibri, produced by MyMax, in: Bonaventura, Maria Antonietta Lozzi (2007): Pompeii Reconstructed, 2007.

[19] Guardasole, *Vesuvio 1270 m*. Guardasole SRL, Napoli, Via Argine, 313, Italia, 2013, Postcard, 40067, Description: 1944 eruptions and present day crater, Collection: LX, Provider: BGS, Entry date: 2013.

[20] M. A. L. Bonaventura, *Pompeii Reconstructed*. ArcheoLibri, 2007, ISBN: 978-88-95512-23-5.

[21] C. G. Newhall and S. Self, "The Volcanic Explosivity Index (VEI): An Estimate of Explosive Magnitude for Historical Volcanism," *JGR*, vol. 87, pp. 1231–1238, 1982.

Spatio Temporal Density Mapping of a Dynamic Phenomenon

Stefan Peters, Liqiu Meng
Department of Cartography
Technische Universität München
München, Germany
e-mail: stefan.peters@bv.tum.de, liqiu.meng@bv.tum.de

***Abstract*— Visualizing density and distribution information is a key support for understanding spatio-temporal phenomena represented by point data. However, the temporal information is not yet adequately handled in existing density map approaches. In this paper, we propose a novel approach for the creation of a 2D density surface - using contour intervals – for dynamic points or phenomena. Our method is based on a rainbow color scheme, which enables the user to visually extract spatio-temporal changes of point density and distribution. Furthermore, we present various possibilities for extending and improving our approach.**

Keywords-spatio temporal density map; rainbow color scheme; visual analysis; dynamic data.

I. Introduction

Visualization helps to investigate and understand complex relationships in a spatial context. Maps account as one of the most powerful visualization forms. They represent geographic information in abstract ways that support the identification of spatial patterns and the interpretation of spatial phenomena.

Furthermore, the visual presentation and analysis of dynamic data and dynamic phenomena is currently a hot research topic [1].

Hence, in today’s society, the need for data abstraction along with the growing amount of available digital geodata is rapidly increasing. One reasonable way of abstracting data is provided by density maps [2].

Density maps can be applied for point data in various fields, for instance, in physical or human geography, geology, medicine, economy or biology [3, 4]. How to present the density for dynamic data/phenomena is, however, not yet adequately addressed.

In this paper, we introduce a novel density mapping approach for spatially and temporally changing data.

In the next section, the state of the art related to density maps, in particular, an overview of approaches considering the dynamics of movement data in the density visualization is given. In the section afterwards, our own approach is described in detail, followed by implementation processes, discussions of the results and a conclusion.

II. Density maps - State of the art

One of the most straight forward ways to visualize point density is a scatter plot or a dot map. Graphic variables for point symbols, such as size, shape, color and transparency, can depend on point’s attribute value. In order to discern the density distribution, these graphic variables can be iteratively adapted to the given map scale, but still the occlusion of neighboring points cannot be always desirably avoided. The density value of each point can be obtained by counting all points within a buffer around the point or within a grid cell the point is located in.

In the following, the density estimation and map principles are shortly presented and the state of the art of density maps with static or dynamic data is given.

A. Kernel Density Estimation (KDE)

The Kernel Density Estimation (KDE) [5] is a classic method widely used to determine densities of individual points that represent a continuous surface. The KDE approach is described in detail in [5, 6, 7]. The standard KDE, a normal distribution function, uses a Gaussian kernel. A certain bandwidth (search radius) is defined for the kernels, located around each point. For each cell of an underlying grid (defined by a certain resolution) a density value is calculated and hence a smooth surface is provided [8]. The bandwidth value strongly influences the density surface [9]. A formula for an optimal bandwidth is proposed by Silverman [6]. Kernel density estimates have been used for cluster detection in various fields, such as crime analysis. Kwan [10] uses geovisualization of activity patterns in space–time and displays the results as a continuous density surface. She applies the density estimation as a method of geovisualization to find patterns in human activities related to other social attributes. Assent, Krieger, Müller, and Seidl [11], Krisp and Špatenková [12], and Maciejewski, et al. [13] investigate the classic kernel density estimation and define it as a visual clustering method. In these works, KDE maps were created in order to visually provide a better overview and insight into the given data.

B. Contour lines and intervals

A common technique to map point densities calculated using KDE are isopleth maps with filled contour intervals. The term “isopleth map” (isopleth = equal in quantity and number), refers to one of two types of isoline maps (also called isarithmic or contour maps). In the first type of isoline maps each contour line indicates a constant rate or ratio derived from the values of a buffer zone or kernel area. In this sense, the continuous density surface is derived from an originally discrete surface. In the other type of isoline maps (commonly referred to “isometric map”), contour lines (isometers) are drawn through points with directly measurable equal value or intensity such as terrain height or temperature [14]. It is assumed that the data collected for enumeration units are part of a smooth, inherently continuous phenomenon [15]. In this paper, we only use contour lines to delimit the intervals (the areas between contour lines). Furthermore, Langford provides a good overview of density surfaces used in Geographic Information Systems (GIS) as choropleth population density maps, population density on grids, population density surfaces, and pseudo-3D population density surfaces [16]. In several works as [3, 4], the KDE concept is adapted for the 3D-Space density mapping of static 3D data.

C. Dynamic data and density information

In the last years, we published several papers related to visual analysis of moving objects with a focus on the representation of dynamic lightning data [17, 18, 19]. In these works, different visualization methods were proposed for moving lightning point data. However, density surfaces were not yet taken into account.

In the following sub-sections, an overview is given about existing works related to density maps of dynamic points.

1) KDE for dynamic points

A straightforward way of visualizing the density of dynamic points would be a sequence of density surfaces (one per time interval) as reported in [20]. The change of the density in time could be better discernable by means of an animation of these density maps. Another option would be to use an individual Kernel density map with a unique color scheme that fills the areas between the contour lines. Due to the movement of points (as for example swarms), however, the tinted intervals may spatially overlap and make the map reading a difficult endeavor.

2) Dual KDE

Jansenberger and Staufer-Steinocher [21] analyzed two different point data sets recorded within the same area, but at two different moments of time. The authors suggest a Dual-KDE approach, which results in a map illustrating the spatio-temporal density difference of the two datasets. The absolute difference is used, that is, the absolute density of the second point data set subtracted from that of the first point data set.

3) DKDE

The approach called Directed Kernel Density Estimation (DKDE) that is able to take the dynamics of moving points inside density maps into account was suggested in previous works [22, 23, 24, 25, 26]. The DKDE is applicable for discrete moving points and it considers two moments of time. Instead of an upright kernel as in the KDE method, a titled kernel is used. The tilt depends on the movement direction vector of the respective point. The resulting DKDE-map shows the so-called “ripples”, which can be interpreted as an indicator for the movement direction and density change of points that are located closely to each other with very similar movement speeds and directions. These ripples are visible among overlapped contour lines. The tinted contour intervals do not contain the information about movement or density change.

4) 3D density map using space time cube

Nakaya and Yano [27] suggest a method using a space–time cube to visually explore the spatio-temporal density distribution of crime data in an interactive 3D GIS. In order to investigate the dynamics and density change, an interactive use within a 3D environment is essential.

5) KDE for trajectories

In a comprehensive review of the existing visual analysis [1], methods, tools and concepts of discrete objects point data were introduced. A section is dedicated to continuous density surfaces (fields) derived from trajectories or from point-related attributes. Density maps of moving objects were created on the basis of aggregated points of trajectories. A trajectory is understood as a function of time or a path left by a moving object in space. Moving objects can be confined within a network (such as cars along streets of a traffic network) or float freely over a region (boats) or in space (airplanes). Spatio-temporal density maps of trajectories were investigated in [28, 29, 30, 31]. In these approaches, the KDE method is adapted to trajectories as a function of moving velocity and direction. The resulting density maps can reveal simultaneously large-scope patterns and fine features of the trajectories. This mapping idea was extended to the 3D space in [29] where the trajectory densities are visualized inside a space-time cube.

Another possibility of displaying density information of trajectories is to use derived discrete grid cells, whereby each cell color refers to the amount of trajectories passing through the cell [32, 33].

D. Research questions

In the existing 2D density maps based on KDE, the time is either frozen on a certain moment or confined within a certain time interval. Consequently, the resulting contour lines do not carry information of temporal changes. Although various approaches for density visualization of trajectories have been investigated, an appropriate method for 2D density maps of moving point clouds is still missing. Can the dynamics of spatially extended phenomena - represented by points - be adequately expressed in a single contour map? This research question remains unsolved. In the following sections, we will tackle this question and introduce a new approach termed Spatio-Temporal Density Mapping or STDmapping.

III. METHODOLOGY

A. Test data

We used lightning points recorded by LINET, a lightning detection network [34], as the test data set. It contains altogether 7100 detected lightnings in the region of Upper Bavaria (47°N–49°N Latitude and 10°E–12,5°E Longitude) on 22.07.2010 between 2pm and 9pm. Each point is encoded with its geographic coordinates (longitude, latitude) as well as the exact lightning occurrence time. The recorded height information is not considered within our approach.

Figure 1 illustrates the provided lightning point coordinates projected onto a plane surface a using black dots.

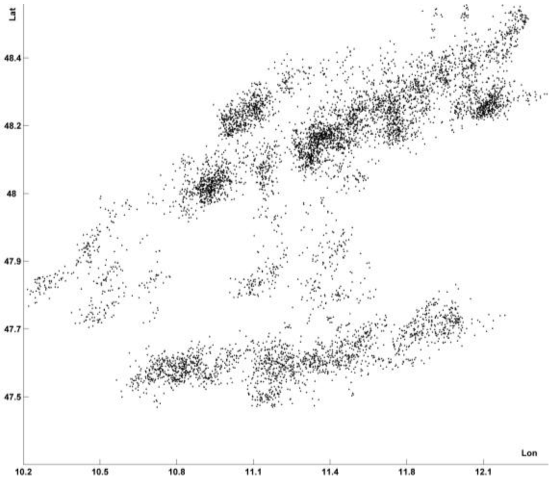


Figure 1. Initial test data set.

Visual analysis methods for these lightning points, which represent the moving phenomena of a thunderstorm, were published in [17, 18, 19].

B. Initial situation and basic idea of the new approach

First of all, we applied KDE using Silverman's formula for calculating an optimal kernel bandwidth [6] to the given test data set. On top of the resulting density contour map, we overlay the initial points, which change colors whenever a time interval has been crossed. In doing so, we used a time interval of 10 minutes starting at 2 pm for the temporal clustering. Additionally, we applied a buffer threshold of 6 km for the spatial clustering.

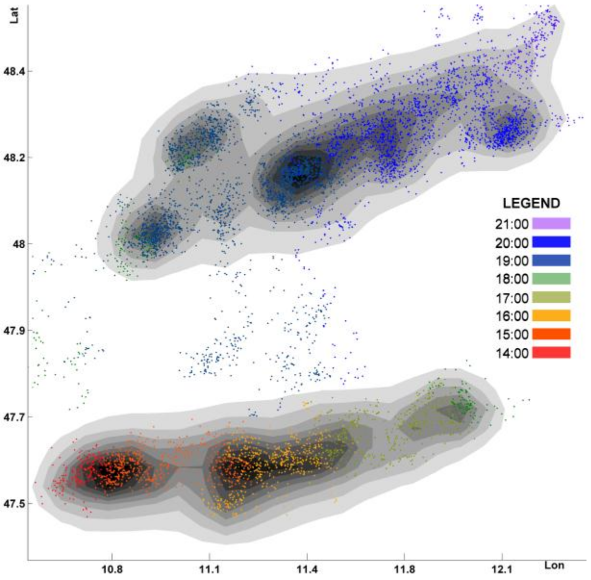


Figure 2. KDE map and clustered points.

Details of the temporal and spatial clustering including explanations for thresholds can be found in [17, 18, 19].

The resulting map is shown in Figure 2. The density layers in grey tones in Figure 2 do not bear any temporal information. But, the overlaid lightning points are segmented and colored according to the different time intervals, thus reveal the dynamic changes. In Figure 2 two moving lightning clusters are perceivable within the test area. Their geographic and temporal locations are apart from each other with one formed lower left starting around 2pm and the other upper left occurring around 6pm. The initially lower left point cluster is moving north-eastwards from 2pm (red dots) to 6pm (green dots). The initially upper left cluster is also moving north-eastwards from 6pm (green dots) to 9pm (purple dots).

C. Workflow

In Figure 3, an overall workflow of our approach is illustrated.

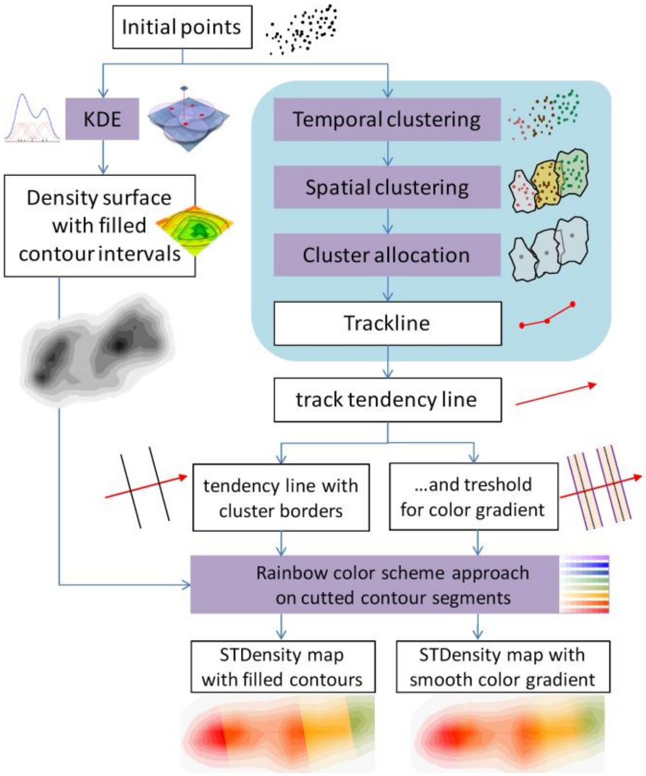


Figure 3. Workflow of STDmapping of lightning data.

Initial lightning point data are explained in Section III.A. As described in Section III.B, first of all a density contour map using KDE is created. Additionally, the given point data set is temporally and spatially clustered. In the next step, the overlapping clusters (in case they are temporally successive) are detected and allocated towards independent tracks. Cluster centroids are embedded in the trajectories. A detailed description about these steps can be found in [17]. A linear approximation of each track trajectory results in a tendency line, which represents the average moving direction of the point cluster. The linear approximation can be based either only on the cluster centroids or on the entire point data sets of a track.

Based on this new approach, we have on the one hand the density surfaces represented by layered tints between neighboring contour lines and on the other hand we have the tendency line with either abrupt or smooth transition at borders of temporal clusters. This temporal border is a line that lies perpendicular to the tendency line passing through the average locations of all points within 10 minutes before and after a full hour. If the phenomenon is moving, all points between the “1 pm line” and the “2 pm line” are grouped into the temporal interval “1 pm - 2 pm”.

The question being addressed now is how we can incorporate the dynamics inside the density map. The idea is to divide the tendency line into temporal parts, which in turn guide the segmentation of the density surface. Different surface segments carry different color hues. Within the same surface segment, the color hue remains the same but its intensity varies with the change of density.

In our approach, we adopted the “rainbow color scheme”, which is essentially the visible and continuous electromagnetic spectrum. Its main color hues transit from red, orange, yellow, green, blue to violet. The spectrum can be divided into an arbitrary number of intervals. Users may easily anticipate and comprehend the color transitions. In our approach, we assign each time interval to a certain color hue – the medium color of the rainbow subinterval. Figure 4 illustrates 8 different rainbow color hues with each being displayed in up to 9 different color intensities from light to dark. For instance, the red color scheme refers to the time from 1 to 2 pm and contains 9 different red tones, which are related to 9 different density values/ value intervals. We split the entire time of our dynamic dataset into equal time intervals. Interval size can be determined based on the user’s interest.

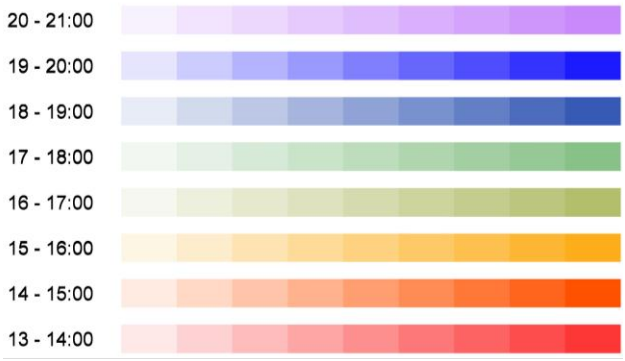


Figure 4. Rainbow color scheme.

With regard to the division of density surface by means of the temporal tendency lines, we introduce the perpendicular line to each tendency line as the temporal border between the two neighboring time intervals of the underlying KDE map. The color transition between two temporal segments can be either abrupt or smooth. In case of smooth temporal borders between temporal KDE segments, a defined threshold for the smooth color transition is set. The threshold refers to a certain time before and after the abrupt temporal borders. That leads to two parallel border lines – one on the left hand site of the abrupt border line and one at the right hand site of the abrupt border line. The distance (time) between each smooth border line and the respective abrupt border line is equal and variable.

IV. RESULTS AND DISCUSSION

As shown in Figure 5 Figure 5. the density visualization option (KDE with dynamic points) described in Section II.C.1) is applied to our test dataset. For each spatio-temporal cluster, a segment of density map with layered tints was produced.

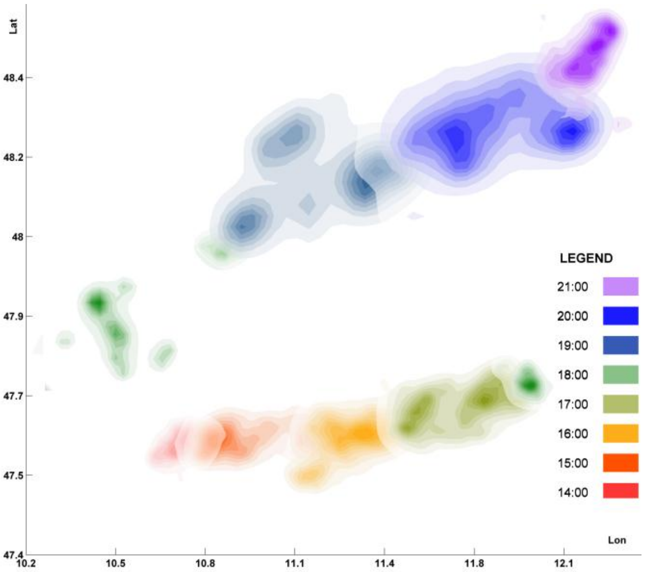


Figure 5. Segmented KDE in one map.

However, the results in Figure 5 are not satisfying due to map overlays and occlusion – even if transparency is applied. It leads to a loss of the overall density information.

Applying the new approach following the workflow in Figure 3, we created two different output maps:

A. *STDmap with abpruptcolor transition*

Figure 6 presents a STDmap with abrupt color transition. Obviously the entire density information is kept while temporal information (and zhus information about phenomena dynamics: speed and moving direction) is the added value: Both lightning clusters are moving north-eastwards and in particular around 8 pm the upper cluster is moving faster than at any other time, while points in the blue colored contours are less dense and more distributed in southwest-northeast direction.

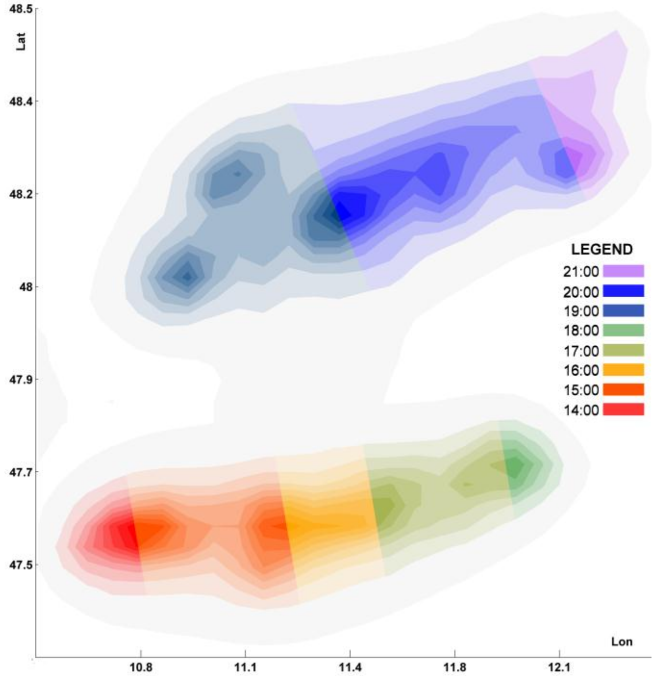


Figure 6. STDmap with abrupt color transition.

The clear-cut temporal cluster borders reveal another advantage: Density information in layered tints within each specific time interval is clearly visible and separable from neighboring segments.

B. Smooth STDmap

Figure 7 presents the STDmap with smooth color transition.

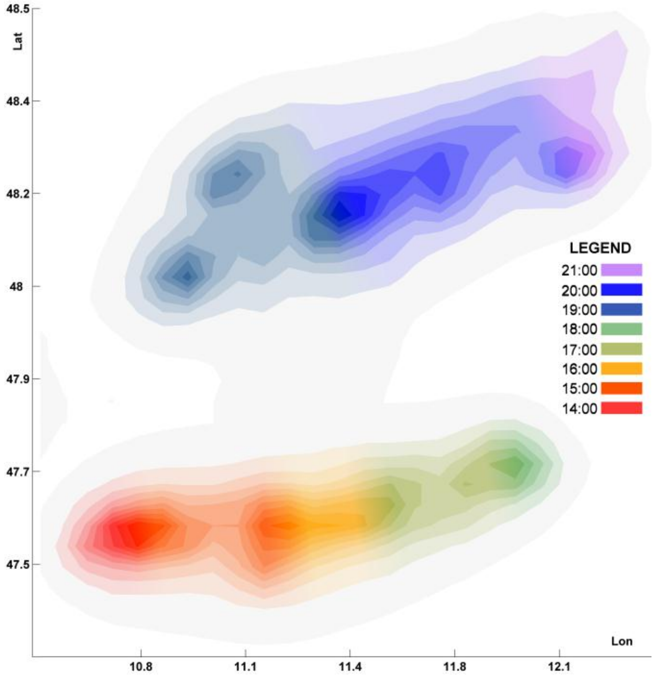


Figure 7. Smooth STDmap.

The smooth color transitions between neighboring segments are closer to the reality and correspond better to the visual perception: lightning points occurring for instance some minutes after 2 pm can be located inside the 1-2 pm segment and points appearing some minutes before 2 pm might be placed inside the 2-3 pm segment. With the help of an adaptive slider tool, the smoothing effect can be set for either a small time interval (e.g., 13:55 – 14:05) or a large one (maximum smoothing interval: half of the time interval left and right of the temporal border, e.g., 13:30 – 14:30). In our case we used a threshold of 20 minutes (10 minutes before and after each abrupt border line). For an easy comprehension, we suggest to limit the number of colors (time intervals) to no more than about 10. In case of very extensive temporal range, the brightness of the same tone within the same interval can be adopted. For instance, 24 hours can be cut into six by four hours intervals. With each four hour interval four different brightness of the same tone can be used.

C. Comparison with existing approaches

The 3D density space time cube suggested by Nakaya and Yano [27] is a comparable approach, where a series of time steps is taken into account within density information visualization with the aim to illustrate the change of point density in time. However, the density changes in time can only be explored by using interactive tools such as panning, zooming and rotating of the space time cube. If a cluster of interest is surrounded by other clusters, it can be hardly explored. Our approach has overcome this drawback by storing and presenting temporal information in different colors in a STDmap (in 2D).

V. CONCLUSION AND FUTURE WORK

In the existing approaches for visualization of dynamic phenomena represented by moving point datasets, temporal information is not yet adequately handed. This research gives a try to close the gap by incorporating and visualizing the temporal change of point cluster in a 2D density map. Our approach is termed as STDmapping according to which a density surface of layered tints can be divided into different temporal segments. Each segment is then visualized by a color hue with varying intensities. The resulted STDmaps contain not only the information about the spatial density distribution, but also the changes in time about moving direction and speed of dynamic point clusters. Therefore, they can support the pattern detection/extraction of spatio-temporal phenomena without having to activate interactive tools.

In future work, we will investigate the relation between the characteristics of initial data (density, distribution, spatio-temporal change of point coordinates) and their modeling parameters (movement tendency, time interval, boundary lines) with the purpose to describe the dynamic phenomena with minimum information loss or distortion for the subsequent visualization and use of STDmaps. Furthermore, an adaption of our approach for 3D point data is also possible.

Moreover, an interesting relative topic could be the dynamic mapping for sensor-based systems: in this case, the contour line must be computed from values regularly sent by sensors (e.g., temperature data).

ACKNOWLEDGMENT

The authors gratefully acknowledge Nowcast Company for providing lightning test dataset and the support of the Graduate Center Civil Geo and Environmental Engineering at Technische Universität München, Germany.

REFERENCES

[1] N. Andrienko and G. Andrienko, "Visual analytics of movement: An overview of methods, tools and procedures," *Information Visualization,* vol. 12, 2013, pp. 3-24.

[2] W. A. Mackaness, A. Ruas, and L. T. Sarjakoski, *Generalisation of geographic information: cartographic modelling and applications*. Amsterdam, The Netherlands: Elsevier Science, 2007.

[3] K. Romanenko, D. Xiao, and B. J. Balcom, "Velocity field measurements in sedimentary rock cores by magnetization prepared 3D SPRITE," *Journal of Magnetic Resonance,* 2012, pp. 120-128.

[4] S. Stoilova-McPhie, B. O. Villoutreix, K. Mertens, G. Kemball-Cook, and A. Holzenburg, "3-Dimensional structure of membrane-bound coagulation factor VIII: modeling of the factor VIII heterodimer within a 3-dimensional density map derived by electron crystallography," *Blood,* vol. 99, 2002, pp. 1215-1223.

[5] J. W. Tukey, *Exploratory data analysis*: Addison-Wesley, 1977.

[6] B. W. Silverman, *Density estimation for statistics and data analysis* vol. 26: CRC press, 1986.

[7] N. Cressie, "Statistics for spatial data," *Terra Nova,* vol. 4, 1992, pp. 613-617.

[8] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization* vol. 383: Wiley. com, 2009.

[9] D. O'Sullivan and D. J. Unwin, *Geographic information analysis*: John Wiley & Sons, 2003.

[10] M.-P. Kwan, "Geovisualisation of activity-travel patterns using 3D geographical information systems," in *10th international conference on travel behaviour research (pp. pages pending), Lucerne*, 2003, pp. 185-203.

[11] I. Assent, R. Krieger, E. Müller, and T. Seidl, "VISA: visual subspace clustering analysis," *ACM SIGKDD Explorations Newsletter,* vol. 9, 2007, pp. 5-12.

[12] J. M. Krisp and O. Špatenková, "Kernel density estimations for visual analysis of emergency response data," in *Geographic Information and Cartography for Risk and Crisis Management*, ed: Springer, 2010, pp. 395-408.

[13] R. Maciejewski, *et al.*, "A visual analytics approach to understanding spatiotemporal hotspots," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 16, 2010, pp. 205-220.

[14] C. F. Schmid and E. H. MacCannell, "Basic Problems, Techniques, and Theory of Isopleth Mapping*," *Journal of the American Statistical Association,* vol. 50, 1955, pp. 220-239.

[15] T. A. Slocum, R. B. McMaster, F. C. Kessler, and H. H. Howard, *Thematic cartography and geovisualization*: Pearson Prentice Hall Upper Saddle River, NJ, 2009.

[16] M. Langford and D. J. Unwin, "Generating and mapping population density surfaces within a geographical information system," *Cartographic Journal, The,* vol. 31, 1994, pp. 21-26.

[17] S. Peters, H.-D. Betz, and L. Meng, "Visual Analysis of Lightning Data Using Space–Time-Cube," in *Cartography from Pole to Pole*, ed: Springer, 2014, pp. 165-176.

[18] S. Peters and L. Meng, "Visual Analysis for Nowcasting of Multidimensional Lightning Data," *ISPRS International Journal of Geo-Information,* vol. 2, 2013, pp. 817-836.

[19] S. Peters, L. Meng, and H. D. Betz, "Analytics approach for Lightning data analysis and cell nowcasting," in *EGU General Assembly Conference Abstracts*, 2013, pp. 32-3.

[20] Z. Zhang, *et al.*, "Nonparametric evaluation of dynamic disease risk: A spatio-temporal kernel approach," *PloS one,* vol. 6, 2011, p. e17381.

[21] E. M. Jansenberger and P. Staufer-Steinnocher, "Dual Kernel density estimation as a method for describing spatio-temporal changes in the upper austrian food retailing market," in *7th AGILE Conference on Geographic Information Science*, 2004, pp. 551-558.

[22] J. M. Krisp and S. Peters, "Directed kernel density estimation (DKDE) for time series visualization," *Annals of GIS,* vol. 17, 2011, 2011, pp. 155-162.

[23] J. M. Krisp, S. Peters, and F. Burkert, "Visualizing Crowd Movement Patterns Using a Directed Kernel Density Estimation," in *Earth Observation of Global Changes (EOGC)*, Munich, Germany, 2013, pp. 255-268.

[24] J. M. Krisp, S. Peters, C. E. Murphy, and H. Fan, "Visual Bandwidth Selection for Kernel Density Maps," *Photogrammetrie - Fernerkundung - Geoinformation*, vol. 2009, 2009/11/01/, 2009, pp. 445-454.

[25] J. M. Krisp, S. Peters, and M. Mustafa, "Application of an Adaptive and Directed Kernel Density Estimation (AD-KDE) for the Visual Analysis of Traffic Data," in *GeoViz2011*, Hamburg, Germany, 2011, pp. 1-1.

[26] S. Peters and J. M. Krisp, "Density calculation for moving points," in *13th AGILE International Conference on Geographic Information Science*, Guimaraes, Portugal, 2010, pp. 10-14.

[27] T. Nakaya and K. Yano, "Visualising Crime Clusters in a Space time Cube: An Exploratory Data analysis Approach Using Space time Kernel Density Estimation and Scan Statistics," *Transactions in GIS,* vol. 14, 2010, pp. 223-239.

[28] MOVE. *WG 4: Visual Analytics for Movement and Cognitive Issues (retrieved: January, 2014)*. Available: http://www.move-cost.info/

[29] U. Demšar and K. Virrantaus, "Space–time density of trajectories: exploring spatio-temporal patterns in movement data," *International Journal of Geographical Information Science,* vol. 24, 2010, pp. 1527-1542.

[30] R. Scheepens, *et al.*, "Composite density maps for multivariate trajectories," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 17, 2011, pp. 2518-2527.

[31] N. Willems, H. Van De Wetering, and J. J. Van Wijk, "Visualization of vessel movements," in *Computer Graphics Forum*, 2009, pp. 959-966.

[32] P. Forer and O. Huisman, "Space, time and sequencing: Substitution at the physical/virtual interface," *Information, Place, and Cyberspace: Issues in Accessibility,* 2000, pp. 73-90.

[33] G. Andrienko and N. Andrienko, "A general framework for using aggregation in visual exploration of movement data," *Cartographic Journal, The,* vol. 47, 2002, pp. 22-40.

[34] H. D. Betz, K. Schmidt, and W. P. Oettinger, "LINET – An International VLF/LF Lightning Detection Network in Europe," in *Lightning: Principles, Instruments and Applications*, H. D. Betz, U. Schumann, and P. Laroche, Eds., ed Dordrecht: Springer Netherlands, 2009, pp. 115-140.

Comparison Between Drainage Network Extracted From Elevation and Surface Models

João Ricardo de Freitas Oliveira, Jussara de Oliveira Ortiz , Sergio Rosim
Image Processing Division
National Institute for Space Research - INPE
São José dos Campos, Brazil
{joao, jussara, sergio}@dpi.inpe.br

***Abstract*—This paper presents a possible procedure to identify critical regions extracting drainage networks from surface model. Qualitative comparison between drainage network extraction from surface and elevation models, both representing the relief, was done. This comparison highlights the differences between drainages extracted from both models and it shows the same critical patterns. For the study area, the radar data was obtained from airborne SAR AES-2 (AeroSensing) with p-band and x-band sensors. Both the elevation model (p-band) and the surface model (x-band) have 2.5m of horizontal resolution. The elevation model represents the actual relief of the land surface, while the surface model is influenced by the coverage of the Earth's surface. This model may show problems in regions with forest, because the canopy of trees forms the relief. Deforestation also causes errors in the drainage representation, leading to spurious drainage formation. To identify where differences occur, a remote sensing image was used. This image was classified to identify forest regions and places with deforestation occurrence. The drainages were superimposed over the classified image to contextualize the critical areas. The remote sensing image was obtained from the Resourcesat-1 satellite, also known as IRS-P6, built by the Indian Space Research Organization, porting the LISS 3 camera operating in three spectral bands (0.52-0.59µm; 0.62-0.68µm; 0.77-0.86µm), yielding 23.5m of horizontal resolution. The *Height Above the Nearest Drainage* (HAND) parameter, a useful terrain descriptor, was used to find the critical areas in the surface model. TerraHidro, a hydrological modeling system, was used to extract the drainage networks.**

Keywords-drainage network; surface model; elevation model.

I. CONCEPTUALIZATION

Drainage network is basic information for studies involving water resources. As a consequence, drainage network extraction must be very precise. Otherwise, the results can lead to wrong decisions. To reach this aim two factors are important: a good drainage extraction method and a good quality Digital Elevation Model (DEM).

If the drainage extraction method is not adequate, unrealistic local features can appear in drainage representation. These features cause wrong representation of the drainage paths, quantitative errors in path length or watershed areas, and other errors. Among the methods to extract adequate drainage network, the Priority First Search (PFS) method was chosen [1] to extract the drainage network with the TerraHidro system.

In relation to data quality, representing the surface of a geographical region, there are two problems to consider: spurious pits and surface versus elevation models.

The spurious pits are created in the generation of the DEM, whether from isolines or image pairs, such as radar images. The interpolation [2], stereoscopic [3], or interferometric [4] processes are responsible to create DEM spurious pits.

The other aspect regards the relief representation data type. There are two data types that represent DEM: one is the elevation model representing the terrain relief, and the other is the surface model, where the altitude is the land cover, for example, the canopy of trees in a forest. As this last type of data model is much more available, it is desirable to identify problems in the generation of drainage networks by using this type of data set.

It is difficult to the user to identify the drainage network positioning errors when he has only the surface model. This work aims to propose a method to determine the most critical error areas. The user can use this information to seek auxiliary data to help him in the manual edition task or, at least, to realize that the results in those regions can be out of acceptable quality range.

This work also shows the differences in drainage network representation when extracted from elevation and surface models to highlight the differences between both drainages and to show that the bigger differences are situated in the most critical areas. To identify the differences, a remote sensing image from Resourcesat-1 [5] was classified and used to understand the problem areas.

Section II describes the relief data models. Section III describes the study region and used data sets. Section IV presents the processing steps adopted. Section V presents the discrepancy analysis of the drainage networks. Section VI proposes a method to determinate the critical areas and makes evaluation of the results obtained. Section VII presents the concluding remarks.

II. RELIEF DATA MODELS

There are two relief data sets to download for free. The first is the Shuttle Radar Topography Mission (SRTM), available at the horizontal resolution of 90 meters [6]. The SRTM data set was generated from images acquired by radar

ISBN: 978-1-61208-326-1

technology. The other data is the ASTER Global Digital Elevation Model (ASTER GDEM) [7] with horizontal resolution of 30 meters from optical images.

Both data sets are surface model representations. As they show land cover, the relief is not well represented everywhere. In this way, a clearing created anywhere in the forest will cause problems in the determination of the drainage network. In large areas, tree tops may be a fair and acceptable relief representation. Deforestation, many trees of different sizes, roads and sudden changing in vegetation type, all these introduce errors in the subjacent relief information, introducing artifacts in the data, lending to an inexact drainage extraction.

Another problem is related to the generation of surface models from texture images. In the SRTM case, texture images are converted to surface model by a procedure called interferometry that creates flat areas in large water regions such as rivers and lakes. In the ASTER case, the images are opticals and the procedure for generating the surface model is called stereoscopy. Optical images suffer direct interference of cloud cover, and where it occurs, the surface values are estimated.

This work shows the drainage networks extracted from surface and elevation models for the same region, with radar technology. The aim is to point out important differences between the drainage networks extracted using both models, and to indicate some way to determine the critical regions prone to error occurrence, using only the surface model data.

III. GEOGRAPHICAL REGION AND USED DATA SETS

The study area of this work is located in the Brazilian Amazon region, (Brazil in green), Pará State (yellow), between latitudes s03^{0}11’03” and s03^{0}03’57”, and longitudes w55^{0}06’03” and w54^{0}54’00” (white rectangle), as shown below in Fig. 1.

For the study area, the radar data was obtained from airborne SAR AES-2 (AeroSensing) [8] with two sensors: p-band and x-band. Both the elevation model (p-band) and the surface model (x-band) have 2.5m of horizontal resolution, comprising an area of 14km by 22km. Fig. 2 presents both models.

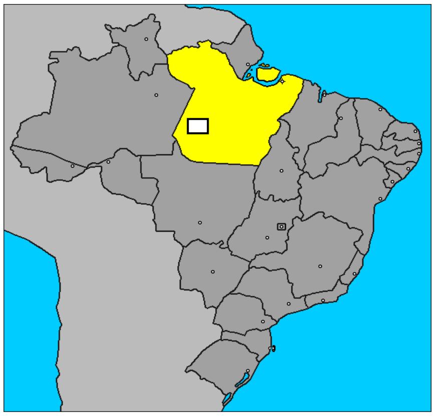


Figure 1. Location of the study area.

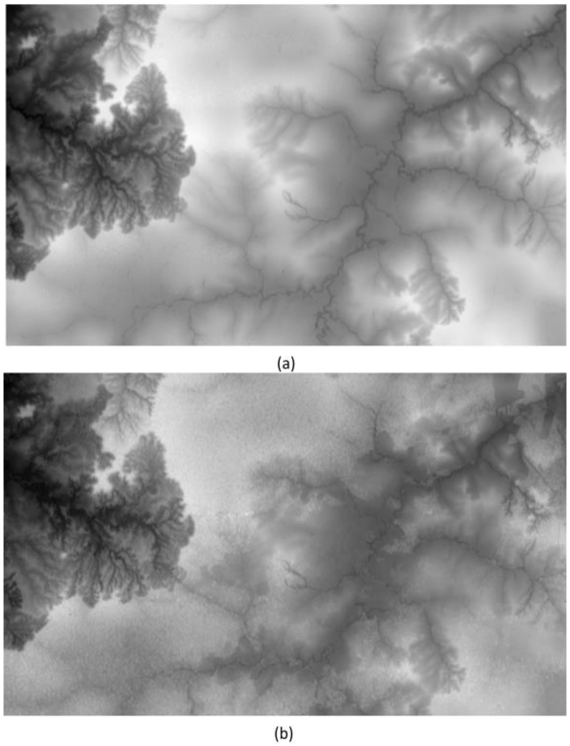


Figure 2. (a) p-band elevation model; (b) x-band surface model.

The other data used was a color image composition from Resourcesat-1, also known as IRS-P6, built by the Indian Space Research Organization, with the LISS 3 camera operating in three spectral bands (0.52-0.59µm; 0.62-0.68µm; 0.77-0.86µm) having 23.5m of horizontal resolution. These images are from 2012. Fig. 3 shows the used color image composition.

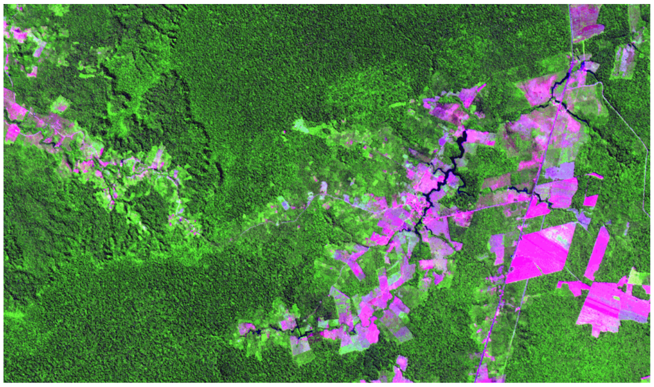


Figure 3. Color image composition: band 1 red, band 2 green and band 3 blue.

The drainage networks were extracted from both radar data and the color image composition was used to verify the differences between the two extracted drainages.

IV. Processing Steps

Drainage was extracted using TerraHidro, a distributed hydrological system for water resource applications [9]. TerraHidro uses a modified PFS method to extract good quality drainage networks [10]. This method eliminates all pits and finds water flow in flat areas. These are the two main problems extracting drainage networks. Fig. 4 presents the drainage extraction for elevation and for surface models.

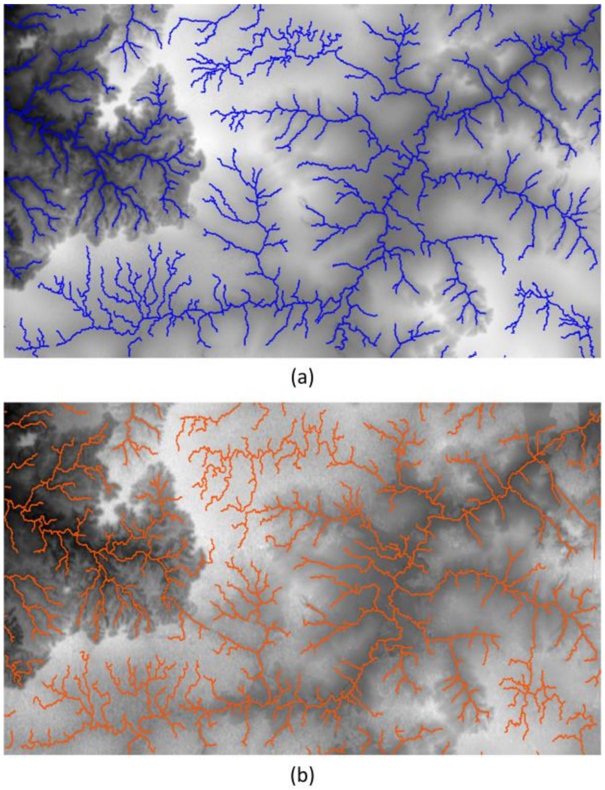


Figure 4. (a) drainage network for elevation model; (b) drainage network for surface model.

Characteristics of land cover can help to identify the differences between both drainages. The images were classified to discriminate areas of water, forest and deforestation. This process was executed in the SPRING system [11] using an automatic classification method, by means of unsupervised classification, using the Isoseg algorithm. Unsupervised classification is a method that works on segmented areas. It examines a large number of unclassified pixels and divides them into a number of classes, based on natural groupings present in the segmented image. The classification process can be described briefly as follows: first, a percentage acceptance threshold is chosen. This threshold is used to calculate the maximum Mahalanobis distance [12] a region created by the segmentation process can be separated from the center of a class and still be considered as belonging to that class. It also determines the number of class clusters detected by the algorithm. Iteratively, the classifier removes all regions with a Mahalanobis distance to any class greater than the acceptance threshold. The user controls the level of details through the acceptance threshold: fewer classes for higher threshold levels or more classes for lower threshold levels. After testing several values, the threshold value of 95% was accepted, which appropriately defined the classes without creating redundancy. Fig. 5 shows the classified image.

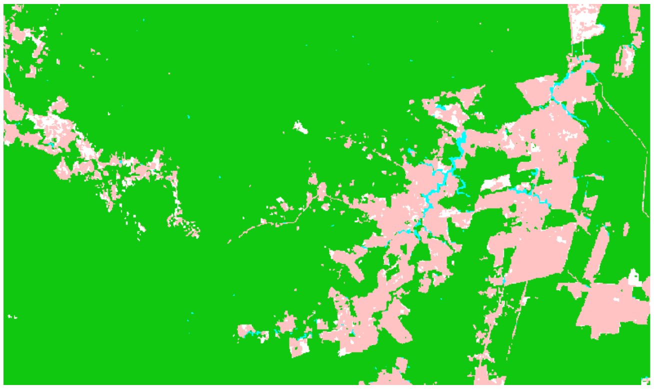


Figure 5. Classified image with forest in green, water in blue, and deforestation in pink.

As the color image composition has 23.5m of horizontal resolution, the drainages must be extracted in the same resolution. TerraHidro uses an upscaling methodology that converts a high resolution drainage network into a low resolution one [13] [14]. In this work, a factor of 9 was used to convert the original grid resolution from 2.5m to 22.5m. As the factor must be an integer, this value yields the best approximation. Fig. 6 shows the drainage networks of the elevation and surface models on top of the classified image.

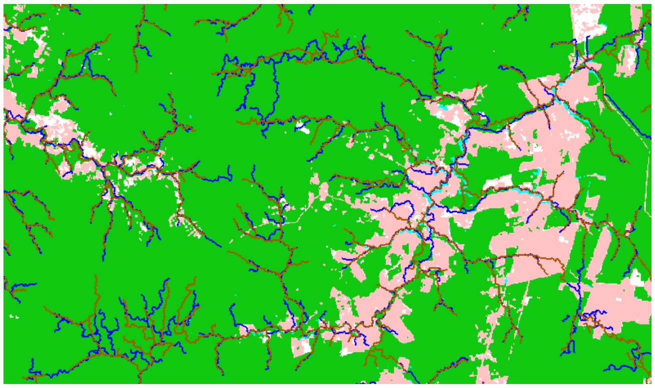


Figure 6. Drainage network for elevation model in orange and drainage network for surface model in blue.

The analysis of the differences between the drainage networks will be qualitative, verifying each drainage portion regarding to the context of the elevation and surface models.

V. Drainage Discrepancy Analysys

The drainage analysis is based on understanding how the drainage changes by using the surface model in place of the elevation model. Two elementary errors occur when using the surface model regarding to the drainage network in

ISBN: 978-1-61208-326-1

deforestation areas and when the relief is represented by canopy of trees. Fig. 7 presents the drainage networks, by using the surface model (blue) and by using the elevation model (white), with the background derived from the classified image, representing in purple the deforestation and forest areas in green.

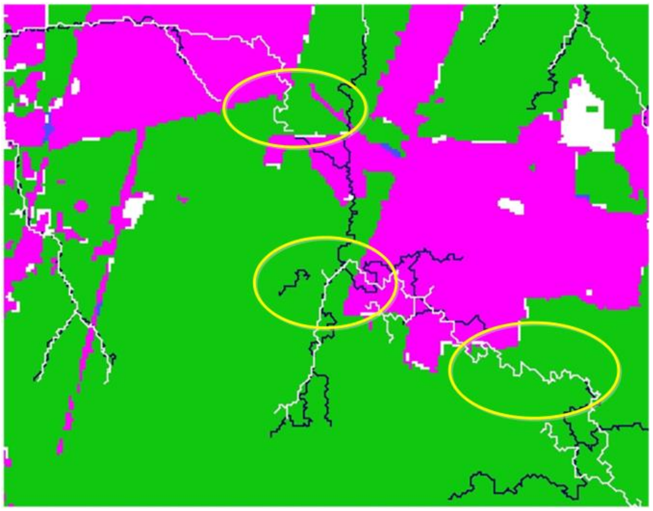


Figure 7. Drainage network for the elevation model in white and drainage network for the surface model in dark blue.

The deforestation creates regions of low elevation with respect to its neighborhood, which is formed by forest. As the elevation is given by the forest (canopy of trees), a path of drainage probably will be erroneously created in the deforested region. Fig. 8 shows the differences between the drainages extracted from elevation and surface models. The yellow ellipses highlight some examples showing significant differences between both drainage networks.

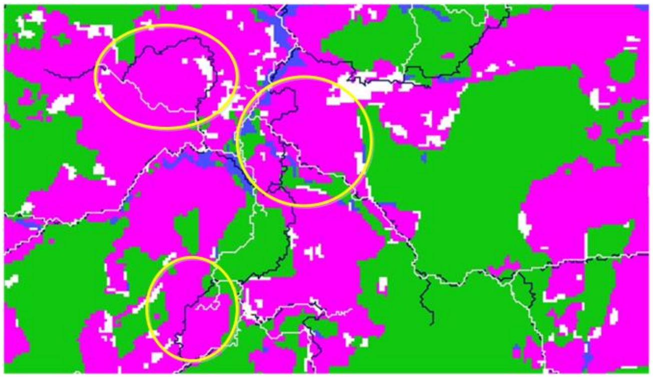


Figure 8. Consequences of using the surface model to extract the drainage network in deforestation areas.

The other problem regards the canopy of trees representing relief. This can causes errors when there are lack of uniformity between the relief formed by the canopy of the trees and the actual terrain topography. Fig. 9 presents some examples of this type of problem.

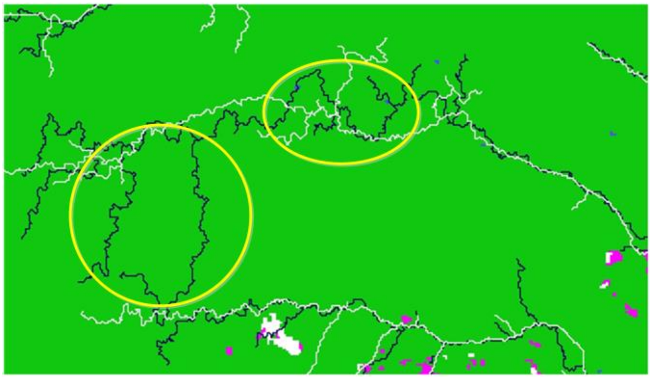


Figure 9. Consequences of using the surface model to extract the drainage network in forest areas.

The yellow ellipses highlight the locations with error occurrence.

VI. Determination and Evaluation of Critical Areas

It was observed that the surface model data, in many cases, varies abruptly in the critical areas, something that is not so common to observe in the elevation model data. As the HAND (Height Above to the Nearest Drainage) [15] terrain descriptor is sensible to drainage changes in regions of sudden terrain variations, it was used as a tentative way to determine critical drainages areas.

HAND calculates, for every cell of the relief regular grid, the altimetry difference between this cell and the nearest cell pertaining to the drainage network, following the local drain directions.

Fig. 10 and Fig. 11 show the resulting DEMs depicting the HAND terrain descriptor for the x- and p-band, respectively, calculated using the TerraHidro System. The x-band drainage is shown in green, and the p-band drainage, in yellow. Critical areas, where the HAND descriptor shows sudden variations in value, are shown inside red ellipses.

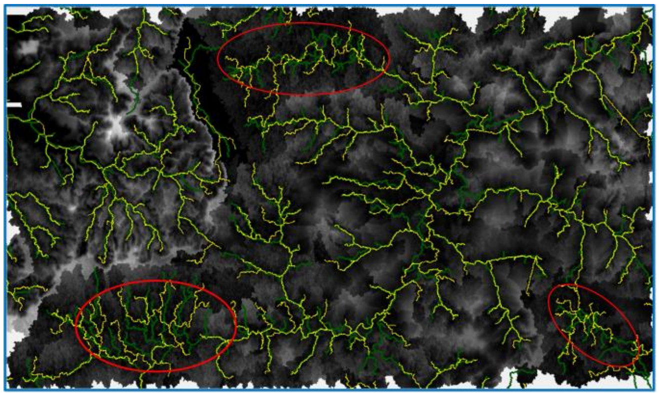


Figure 10. Descriptor HAND for the x-band.

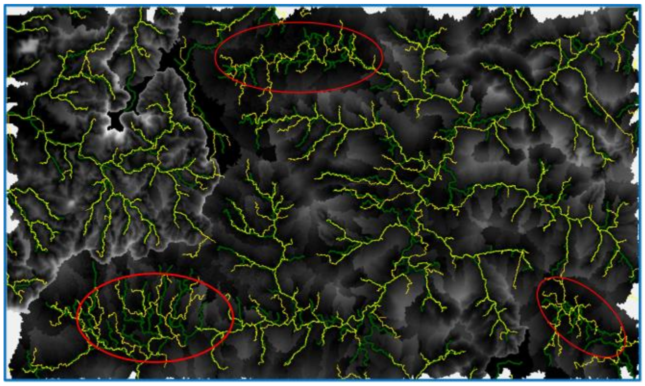


Figure 11. Descriptor HAND for the p-band.

The use of the HAND descriptor can help to spot areas where the drainage network is less precise and may contain errors. It can be useful if an x-band radar data DEM is the only data available.

VII. Conclusion

This article presented the drainage networks differences when they are extracted from the surface model and from the elevation model. Real data was used for both models and two situations causing problems, deforestation and forest areas, were shown.

Drainage networks extracted from surface models have errors that can significantly impair the quality of the results. Not always, however, it can be easily identified. Use of classified images, identification of deforestation and forest regions, are strategies to make better drainage extraction.

This work also presents a procedure to determine critical areas for extracting drainage network from x-band radar images. A limitation of this procedure is that it shows only the most critical areas. Other critical areas are not found by the suggested procedure.

This procedure can be useful when there is only an x-band DEM available, with no additional information. It signals potential critical areas, being necessary to have more data or information about these areas to correct the extracted drainages. In the worst case, when there are no other data, the drainage network extracted in these areas is not to be trusted.

This issue is important because large databases of reliefs, such as SRTM and ASTER, were created from surface models and are worldwide used.

The development of this work had the financial support of FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), under process 2014/01254-9.

References

[1] R. Jones, “Algorithms for using a DEM for mapping catchment areas of stream sediment samples,” Computers & Geosciences, vol. 28, 2002, pp. 1051–1060.

[2] R. Fencík and M. Vajsáblová, "Parameters of interpolation methods of creation of digital model of landscape," 9th AGILE Conference on Geographic Information Science, 2006, pp. 374-381.

[3] Cheng P. and C. Chaapel, "Automatic DEM generation". http://www.pcigeomatics.com/pdf/GeoInformatics08_WorldView-1_DEM.pdf [retrieved: 03, 2014]

[4] Dungchen E., Chunxia Z., and Minsheng L., "Application of SAR interferometry on DEM generation of the Grove mountains," Photogrammetric Engineering and Remote Sensing, vol. 70, 2004, pp. 1145-1149.

[5] G. Chander, "Overview of the Resourcesat-1 (IRS-P6)". http://calval.cr.usgs.gov/documents/IRSP6.pdf [retrieved: 03, 2014]

[6] SRTM - Shuttle Radar Topography Mission of NASA - National Aeronautics and Space Administration. http://www2.jpl.nasa.gov/srtm/ [retrieved: 03, 2014]

[7] ASTER Global Digital Elevation Model (GDEM). http://www.jspacesystems.or.jp/ersdac/GDEM/E/4.html [retrieved: 03, 2014]

[8] J. Moreira, "Design of an airborne interferometric SAR for high precision DEM generation," International Archives of Photogrammetry and Remote Sensing, vol. XXXI, part B2, 1996, pp. 256-260.

[9] S. Rosim, J. R. F. Oliveira, A. C. Jardim, L. M. Namikawa, and C. D. Rennó, “TerraHidro: A distributed hydrology modelling system with high quality drainage extraction,” GEOProcessing 2013: The Fifth International Conference on Advanced Geographic Information Systems, Applications, and Services, 2013, pp. 161-167.

[10] W. Collischonn, D. C. Buarque, A. R. Paz, C. A. B. Mendes, and F. M. Fan, “Impact of pit removal methods on DEM derived drainage lines in flat regions”, AWRA 2010 Spring Specialty Conference, 2010, pp. 1-6.

[11] G. Câmara, R. C. M. Souza, U. M. Freitas, J. Garrido, and F. M. Ii, “SPRING: integrating remote sensing and GIS by object-oriented data modelling,” Computers & Graphics, vol. 20(3), 1996, pp. 395-403.

[12] J. A. Richards, Remote Sensing Digital Image Analysis: an Introduction, 2nd ed., Berlin: Springer-Verlag, 1995.

[13] S. M. Reed, “Deriving flow directions for coarse-resolution (1-4km) gridded hydrologic modeling,” Water Resources Research, vol. 39(9), 1238, 2003, pp. 4.1 – 4.11, doi:10.1029/2003WR001989

[14] A. R. Paz, W. Collischonn, and A. L. L. Silveira,” Improvements in large-scale drainage networks derived from digital elevation models,” Water Resources Research, vol. 42, W08502, 2006, pp. 1-7, doi:10.1029/2005WR004544

[15] C. D. Rennó, A. D. Nobre, L. A. Cuartas, J. V. Soares, M. G. Hodnett, J. Tomasella, and M. J. Waterloo, “HAND, a new terrain descriptor using SRTM-DEM: mapping terra-firme rainforest environments in Amazonia,” Remote Sensing of Environment, v.112, 2008, pp. 3469-3481.

Trading off Accuracy and Computational Efficiency of an Afforestation Site Location Method for Minimizing Sediment Yield in a River Catchment

René Estrella*, Pablo Vanegas†, Dirk Cattrysse‡ and Jos Van Orshoven*

* Division of Forest, Nature and Landscape
Department of Earth and Environmental Sciences
KU Leuven, Leuven, Belgium
Email: {rene.estrellamaldonado, jos.vanorshoven}@ees.kuleuven.be

† Facultad de Ingeniería
Universidad de Cuenca, Cuenca, Ecuador
Email: pablo.vanegas@ucuenca.edu.ec

‡ Centre for Industrial Management / Traffic & Infrastructure
Department of Mechanical Engineering
KU Leuven, Leuven, Belgium
Email: dirk.cattrysse@cib.kuleuven.be

***Abstract*—The Cellular Automata based method for Minimizing Flow (CAMF) aims at selecting, from a rasterized database representing a river catchment, a predefined number of cells that should be afforested in order to minimize the sediment yield of the catchment. To this end, CAMF iteratively ranks cells according to sediment yield reduction, taking into account spatial interaction among cells. It was found during tests that the execution time of CAMF is directly proportional to the database size and the number of cells to be selected. This behavior can become a limiting factor for the applicability of CAMF to high resolution databases that cover large geographical areas. This issue motivated the necessity of exploring simplified CAMF variants that reduce its execution time and preserve the accuracy of its results. For this purpose, a simplified variant called on-site CAMF was devised, implemented and tested. On-site CAMF ranks cells based only on local cell information, i.e., the local sediment reduction that afforestation would produce in a cell, and the cell slope. During tests, on-site CAMF produced virtually the same results as the original version of CAMF in only a small fraction of the execution time. This means that, for these particular tests, spatial interaction did not influence CAMF output, possibly due to the number of cells that were selected, which was small with respect to the full geodatabase size. It is expected that spatial interaction becomes a relevant factor when larger sets of cells are selected.**

Keywords–Site location; Spatial interaction; Sediment yield; Optimization; Afforestation.

I. INTRODUCTION

Soil erosion is a common problem in tropical mountainous regions. In such regions, rainfall typically produces high levels of runoff, which in turn causes the soil to be eroded and, as a consequence, large amounts of sediment are produced, transported and deposited [1]. This often leads to the undesirable result of degraded soil, i.e., soil with severely limited performance in terms of fertility and productivity. A second negative consequence caused by soil erosion occurs when the sediment produced is delivered to the river system of a catchment. This sediment will be partially transported so that it will eventually reach the outlet of the catchment. This process is a critical factor when there exists a dammed reservoir downstream the river, since the sediment input to such infrastructures might produce high costs for sediment removal and a shortening of the reservoir lifespan given the resulting loss of capacity [2]. These factors make the study of sediment flow in mountainous regions crucially important.

One measure that has proven useful to control sediment production is afforestation ([3], [4], [5]), especially when it is technically planned and based on sufficient scientifically sound information. Typically, when planning an afforestation project, several criteria are to be considered simultaneously. Some of these criteria may pertain to the local performance of areas within the study region. This type of criteria are referred to as on-site. An example of on-site criteria is the amount of carbon sequestered both in soil and in biomass. On the other hand, some criteria can be related to the effect that changes in the state of a given area produce in the state of neighboring or even distant areas within the study region. These criteria are classified as off-site. Sediment delivery to the river and sediment yield of a river catchment are examples of this type of criteria. Both on-site as well as off-site criteria allow forest planners to discriminate between suitable and unsuitable alternatives, e.g., selecting sites for afforestation, choosing the species to be planted, or deciding when to harvest the forest.

The term site location for afforestation used throughout this paper refers to determining the exact locations in which trees should be planted. In this specific case, decision alternatives correspond to candidate sites within a river catchment that are available to be afforested. Only areas under agriculture and pasture are considered as candidates for afforestation. A single off-site factor, the amount of sediment at the outlet of the catchment, or sediment yield, was chosen as the decision criterion. Since the study regions are represented by raster datasets, the problem amounts to selecting a subset of cells (pixels) that should be afforested in order to minimize the

ISBN: 978-1-61208-326-1

sediment yield of a river catchment.

A computational iterative method to tackle this problem was proposed in [6]. This method aimed to select, at each iteration, the cell(s) that, in case of being afforested, would produce the maximum reduction in sediment yield. The name Cellular Automata-based method for Minimizing Flow (CAMF) was used to refer to this method. To select a cell or cells at each iteration, CAMF computes the sediment yield reduction that would be produced considering that every candidate cell is afforested separately. This sediment yield reduction values is then used to build a ranking from which the optimal cell(s) is (are) selected.

Some limitations were identified in CAMF. One of these limitations is the fact that scoring cells and building the ranking are relatively expensive procedures in terms of execution time. A second limitation is that there is a high probability that only one cell is selected at each iteration, so that many iterations of CAMF are necessary in order to select the required number of cells. This undesirable combination of repeating many times a computationally expensive procedure might restrict the applicability of CAMF when dealing with high resolution datasets that cover extensive study areas.

This work aimed at providing insights about several aspects of CAMF. First, the performance of CAMF was examined as a function of the size of the database to which it is applied and of the number of cells to be selected. This analysis produced indicators about the applicability of CAMF to large databases, which are frequently found in natural resources management projects. This goal was meant to complement the work reported in [6], where only very small, sample databases were used during tests. The second general aim was to propose a variant of CAMF that addresses its limitations to drastically reduce its execution time while preserving the quality of the results it produces.

Section II introduces the study regions and the corresponding geodatabases that were used during tests. This section also explains CAMF and its on-site variant in detail as well as the performance indicators that were collected during tests. Section III presents and discusses the results produced by CAMF and its on-site variant. Finally, Section IV draws some conclusions and proposes a few points that require further work.

II. Materials and Methods

A. Study regions

Three raster geodatabases were used for testing CAMF. These geodatabases, stored using the ArcInfo ASCII grid format, represent nested river catchments located in the southern Andes of Ecuador using a cell resolution of 30x30 m. The study regions and their corresponding geodatabases are introduced below.

1) The Paute river catchment: The Paute river catchment is located in the southern Andes of Ecuador. Its area is 5055 km^2. Altitudes in this catchment vary between 1591 and 4651 m asl. High sediment production rates have been measured in this catchment in the past [1] and several dammed reservoirs that are part of one of the most important Ecuadorian hydroelectric complexes are located within this catchment. The location in Ecuador and the sediment production of the Paute catchment are shown in Figure 1. The areas under agriculture and pasture in this catchment correspond to a total of 1483 km^2 (around 30% of the full area of the catchment).

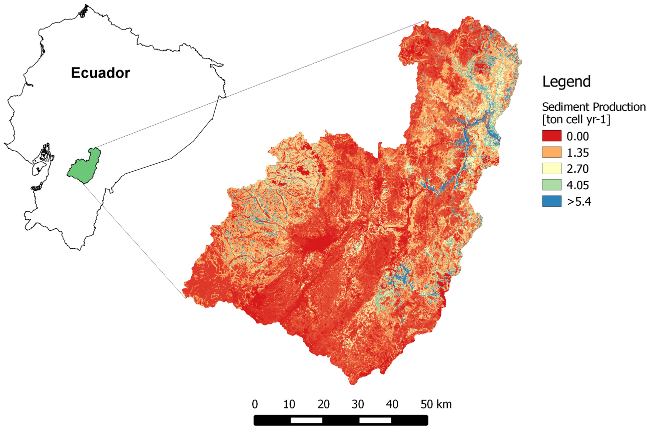


Figure 1. Location and sediment production of the Paute catchment. Cell size is 30x30 m

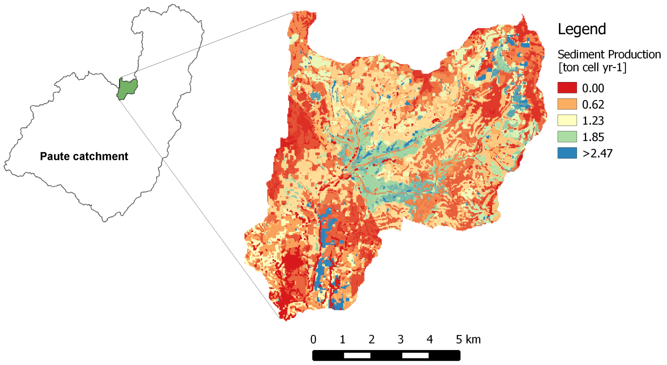


Figure 2. Location and sediment production of the Tabacay catchment. Cell size is 30x30 m

2) The Tabacay river catchment: Tabacay is a subcatchment of the Paute catchment. Its total area amounts to 66.3 km^2. The altitudes are in the range between 2481 and 3732 m asl. The importance of studying sediment production and transport in this catchment is given by the fact that the Tabacay river, besides being a tributary of the Paute river, is used as the source for provision of drinking water to the city of Azogues. Agriculture and pasture cover a region of 24 km^2 in the Tabacay catchment (39% of the total catchment area). Figure 2 depicts the location of Tabacay within the Paute catchment and its sediment production.

3) The Tabacay500 database: The third database used in tests corresponds to a part of the Tabacay catchment, which represents an area of 1.7 km^2 around its outlet. The codename Tabacay500 was chosen for this database because it comprises 500 cells (26% of the full area) that are considered as the initial candidates for afforestation. The location of the region represented by this database within Tabacay and its sediment production are displayed in Figure 3.

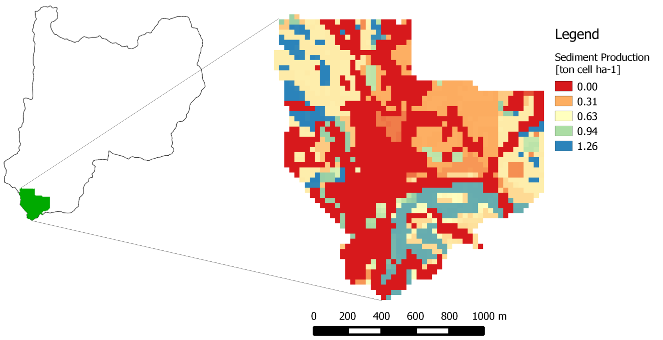


Figure 3. Location and sediment production of the region represented by the Tabacay500 database. Cell size is 30x30 m

B. Cellular Automata based method for Minimizing Flow (CAMF)

[6] introduced a computational iterative method aimed to locate sites that, after afforestation, would result in the minimization of the sediment yield of a river catchment. In [6], the acronym CAMF is used to refer to this method, which is described in the following subsections.

1) Required input data: To execute CAMF the following input data are required:

Sediment production:
: This is a raster dataset containing values about the sediment produced locally in each cell, expressed in $ton\,cell\,yr^{-1}$;

Retention capacity:
: If the amount of sediment in a cell is smaller than its retention capacity, expressed in $ton\,cell\,yr^{-1}$, it is assumed that no sediment leaves that cell;

Saturation threshold:
: The amount of sediment in a cell that exceeds the saturation point, expressed in $ton\,cell\,yr^{-1}$, is assumed to be fully delivered to its steepest downslope neighbor;

Flow factor:
: Raster dataset that indicates the fraction of the amount of sediment in a cell that is delivered to one of its neighbors. This fraction is applied only when the amount of sediment in a cell is in the range between the retention capacity and the saturation threshold;

Flow direction:
: CAMF uses a Single Flow Direction (SFD) dataset based on the Deterministic 8 (D8, [7]) method to determine the flow path that sediment follows within a catchment. The D8 method assumes that flow leaving a cell is delivered only to its steepest downslope neighbor;

Solution size:
: Parameter set by the user of CAMF to indicate how many cells should be selected to be afforested.

Two different versions of each of the first four datasets listed above are required: 1) a dataset representing the initial situation, that is, the catchment under its original land cover; and 2) a dataset that represents the catchment in case every cell was under forest. These two versions of each of the four datasets are used by CAMF to compute the amount of sediment that leaves each cell. The flow direction dataset is used to simulate the transport of the sediment within the catchment. In other words, the flow direction dataset allows to incorporate spatial interaction into CAMF, which in turn permits the involvement of off site criteria, like sediment yield.

2) Workflow: CAMF is an iterative method that comprises the following steps:

1) The sediment accumulation in each cell is computed. Sediment accumulation refers to the sediment locally produced in a cell plus the amount of sediment that it receives from its neighbors;
2) For each candidate cell, the sediment yield reduction that would occur in case that cell is afforested is computed;
3) A ranking of all candidate cells is built based on the sediment yield reduction values computed in the previous step;
4) The cell or cells at the top of the ranking that correspond to the maximum score are selected as part of the solution;
5) The sediment accumulation values are updated for the selected cells and for all the cells that are between each selected cell and the outlet;
6) If the total number of selected cells is less than the solution size, repeat from step 2.

As a first step, an implementation of CAMF as described in [6] and outlined above was produced. This implementation is referred to as ‘original CAMF’. The purpose of implementing and testing original CAMF was threefold. First, to explore the applicability of CAMF to databases that are larger than the ones used in [6]. The second objective was to produce reference values for comparison to the variant of CAMF introduced below. The third objective was to approximate the average number of cells that are selected by CAMF at each iteration.

After studying original CAMF, two issues were pinpointed that can compromise the computational efficiency or even the applicability of this method, namely:

1) At each iteration, CAMF computes the sediment yield reduction that would be produced in case every single candidate cell in the catchment was afforested. Depending on the extent covered and on the resolution of the database, the number of candidate cells can reach several millions. Note that the computation of the sediment yield reduction for a single cell requires simulating the sediment transport from that cell to the outlet. After this has been done for every candidate cell, a sorted list of cells (ranking) is built. It was expected then that the computational time required to build this ranking is relatively high.
2) Once the ranking is built, only the cell or cells at the top of the ranking that correspond exactly to the maximum sediment yield reduction are selected. It is unlikely that many cells correspond exactly to the same sediment yield reduction value. As a consequence, it was expected that only one cell is

selected at each iteration, which would result in a limited use of the computationally expensive ranking mentioned above.

As mentioned above, sediment yield minimization is an example of an off-site criteria. This means that the sediment yield reduction that is produced when a given cell is afforested not only depends on local information, but also on information pertaining to other cells, i.e., the outlet and all cells in the steepest descent path between it and the considered cell. As already explained, computing sediment yield reduction values in CAMF involves the notion of spatial interaction, which is intuitively appropriate, especially for the case of sediment transport in mountainous regions. On the other hand, taking spatial interaction into account is a decision that contributes to a large extent of the computational cost of original CAMF in terms of execution time. This issue is dealt with by the CAMF variant proposed in the following section.

C. On-site CAMF

On-site CAMF aims at avoiding the extra computational cost that considering spatial interaction introduces into CAMF operation. In on-site CAMF, the notion of spatial interaction is simply disregarded and only local (on-site) information is used to rank cells. The motivation of on-site CAMF is based on the claim made by [6], which states that, in general, cells with steep slopes and high local sediment production are selected by CAMF. Based on this conclusion and considering that these two factors correspond to on-site information that was readily available for the study regions, they were chosen as the basis to compute scores that allow to produce a cell ranking in this variant of CAMF. The score assigned by on-site CAMF to a cell was computed using (1).

$$s_i = w_f f_i + w_e e_i \tag{1}$$

where

- s_i is the score assigned to cell i.
- w_f and w_e are user defined parameters that can take values in the range [0, 1] and indicate the relative importance (weight) assigned to each factor, either slope or sediment production, respectively, with $w_f + w_e = 1$.
- f_i is the normalized (scaled to the range [0, 1]) slope of cell i.
- e_i is the normalized change that would be produced in local sediment production when cell i was afforested, that is the difference in sediment production between the initial situation and the afforested situation.

Note that cells with higher values for s_i will be preferred to be selected. Note as well that both slope and local sediment production values do not change during the execution of this method, which means that on-site CAMF is not an iterative method and, therefore, all required cells are selected in a single step.

D. Methodology

1) Performance measures: The experimental phase consisted in several executions of both original and on-site CAMF for the three databases described in Section II-A for solution sizes corresponding to 1, 10, 100 and 1000 cells. During each test, several performance factors were recorded, namely:

Sediment yield reduction:
: Decrease in the sediment yield of a catchment (with respect to the initial situation) when the required number of cells are afforested;

Execution time:
: CPU time necessary to produce the required output;

Number of iterations:
: Number of iterations performed by original CAMF to produce the required output;

Spatial coincidence:
: This is a comparative performance measure applicable only to on-site CAMF. It uses the output (cells selected for afforestation) produced by original CAMF as a reference. Spatial coincidence was computed as $\frac{n_c}{n}$, where n_c is the number of common cells selected by both original and on-site CAMF, and n is the solution size. Therefore, a spatial coincidence of 1 indicates that both methods selected exactly the same set of cells.

2) Parameter values: The different input datasets and parameter values used when executing all versions of CAMF are listed in TABLE I. The values corresponding to retention capacity and saturation threshold were arbitrarily set in such a way that around half of the available sediment under the original land cover leaves the river catchment in a time unit (year).

III. Results and Discussion

A. Original CAMF

The output and performance measures obtained after executing original CAMF are shown in TABLE II. The sediment yield reduction in case the corresponding number of cells are afforested is shown as an absolute value in the second column and as a percentage with respect to the initial sediment yield in the third column. The last column shows the average number of cells that are selected at each iteration.

It is clear from TABLE II that sediment yield reduction values for Tabacay500 and Paute increase almost proportionally with respect to the solution size, which is an indication that at least 100 cells in Tabacay500 and 1000 cells in Paute perform almost equally well when afforested. This is not the case for Tabacay, where such proportionality is evident only when comparing the sediment yield reduction corresponding to solutions sizes of 1 and 10. This proportionality is not present when solutions sizes of 100 and 1000 cells are considered. Except for Tabacay500, execution times seem to increase also in a direct proportion with respect to solution sizes. This is given by the fact that in almost all cases the number of iterations performed by original CAMF is equal to or slightly smaller than the corresponding solution size. This effect is less noticeable for Tabacay500, for which only very short times are

TABLE I. INPUT DATA AND PARAMETER VALUES USED DURING EXPERIMENTATION PHASE

Input/Parameter		Dataset/Value
Sediment production [$ton\, cell\, yr^{-1}$]		Available datasets (Figure 1, 2 and 3)
Retention capacity [$ton\, cell\, yr^{-1}$]	initial	Paute: 0.27, Tabacay: 0.17, Tabacay500: 0.075
	afforested	Paute: 0.54, Tabacay: 0.34, Tabacay500: 0.15
Saturation threshold [$ton\, cell\, yr^{-1}$]	initial	Paute: 0.81, Tabacay: 0.51, Tabacay500: 0.225
	afforested	Paute: 1.08, Tabacay: 0.68, Tabacay500: 0.3
Flow factor [-]	initial	Slope linearly scaled to [0, 1]
	afforested	Initial flow factor divided by 2
Flow direction [-]		Computed from DEM, based on D8 [7]
Solution size		1, 10, 100, 1000

TABLE II. PERFORMANCE MEASURES CORRESPONDING TO ORIGINAL CAMF

Solution size	SYR [$ton\, yr^{-1}$]	% SYR	CPU time [s]	# iterations	Cells/iteration
Tabacay500 (initial SY: 370 $ton\, yr^{-1}$, total cells 1892, candidate cells 500)					
1	0.498	0.1	0.015	1	1.00
10	4.971	1.3	0.046	10	1.00
100	46.665	12.6	0.109	97	1.03
Tabacay (initial SY: 29075 $ton\, yr^{-1}$, total cells 68123, candidate cells 26850)					
1	3.308	0.01	0.234	1	1.00
10	32.171	0.11	1.872	10	1.00
100	199.806	0.69	17.799	99	1.01
1000	924.975	3.18	155.002	927	1.08
Paute (initial SY: 3212203 $ton\, yr^{-1}$, total cells 5616679, candidate cells 1647304)					
1	14.729	0.0005	133.646	1	1.00
10	147.205	0.0046	1311.625	10	1.00
100	1470.557	0.0458	11556.164	87	1.15
1000	14675.398	0.4569	93484.394	701	1.43

required. In this case, internal details of the implementation and even technical aspects related to the way in which the algorithm is executed by the operating system take a higher relative importance with respect to factors pertaining to the method itself, like simulating sediment flow and building the ranking of cells.

Unexpectedly, in all tests involving solution sizes of 100 and 1000, the number of cells selected per iteration by original CAMF is greater than one. This finding indicates that the probability of more than one cell corresponding to exactly the same maximum sediment yield reduction at a given iteration plays a role in practice. This may be an indication that the function applied to compute the amount of sediment leaving a cell and the way in which sediment flow is simulated, play a homogenizing role for the computation of sediment yield reduction values. On the other hand, in all those tests, the average number of cells selected per iteration is still close to one. This characteristic may make CAMF execution times unnecessarily long.

Database size, in terms of number of candidate cells, has a clear impact on execution times. This is explained by the fact that larger database sizes will require more cells to be processed at each iteration. When comparing the execution times obtained for Paute to the corresponding values for Tabacay, a clear proportionality is found. This is not the case when execution times for Tabacay and Tabacay500 are contrasted. This may be the result of (very short) execution times for Tabacay500 being largely influenced by internal, technical aspects of algorithm execution. When considering execution times separately, it can be argued that they start to play a restrictive role for large databases like Paute. Specifically, original CAMF requires more than 3 hours to select 100 cells in Paute, and almost 26 hours for a solution size of 1000 cells. Additionally, it is important to note that the solutions sizes tested are rather limited, considering the full size of the database, especially for the case of Paute.

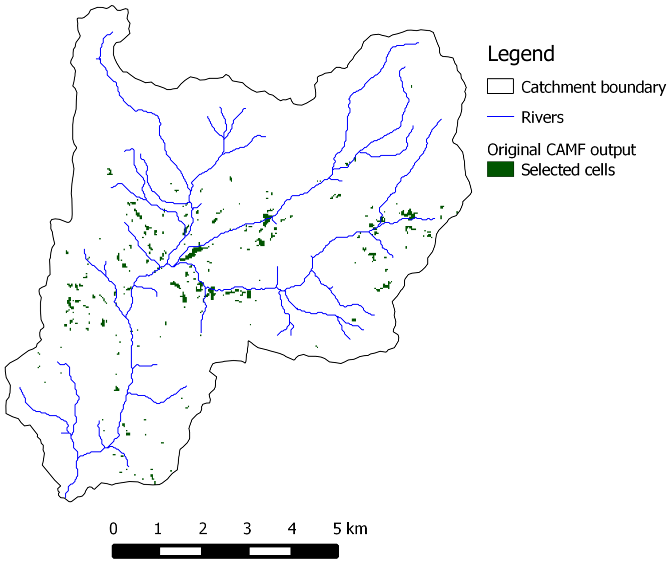


Figure 4. Output of original CAMF for a solution size of 1000 cells in Tabacay

Figure 4 shows the 1000 cells that were selected by original CAMF in Tabacay.

B. On-site CAMF

The first step conducted when applying on-site CAMF was to determine sensible values for the relative importance parameters corresponding to slope and sediment (w_f and w_e in (1)). In this case, a naive trial-and-error approach was used, based on testing different combinations of values for w_f and w_e to score, rank and select cells and assessing the corresponding values of sediment yield reduction produced when the selected cells were afforested. The tested parameter values and the resulting sediment yield reduction values are listed in TABLE III.

The values in columns 2 to 6 of TABLE III show the ratio between the sediment yield reduction produced by on-site CAMF when the relative importance parameters were set to the values indicated in the headers and the sediment yield reduction value produced by original CAMF for the corresponding database and solution size. From this we can conclude that setting $w_f = 0.01$ and $w_e = 0.99$ produces the best results from among the tested combinations. This means that slope plays a very limited role in cell selection in CAMF, whereas local sediment reduction appears as the most relevant factor in this regard. These values were used in all tests performed with on-site CAMF to produce the performance indicators listed in TABLE IV. Column 'SYR fraction' shows the ratio between the absolute sediment yield reduction resulting from on-site CAMF and original CAMF. Similarly, 'CPU time fraction' lists the ratio between the execution time of on-site CAMF with respect to original CAMF.

It can be seen from TABLE IV that on-site CAMF produces practically the same results as original CAMF. A first interpretation of these results is that spatial interaction does not play a role for the combination of databases and parameter values used during the tests. Considering the values set for the relative importance parameters (w_f and w_e), it can be argued that the local sediment reduction information, that is, the amount in which sediment production would decrease in every cell when afforested, is virtually the only factor that is determining which cells are selected.

Stating that spatial interaction does not play a role when sediment transport simulation in particular, and off-site criteria in general, are involved may seem counter intuitive. However, this finding can be supported by the fact that relatively limited solution sizes were used during the tests, especially for the cases of Tabacay and Paute. When a limited number of cells are to be selected from a large number of candidate cells, it can occur that most selected cells in fact does not interact with each other, which means that they do not share a meaningful segment of their path to the outlet and therefore, changes in the state of one cell do not affect the state of other selected cells. It can be claimed that the river may act as an element that produces interaction among cells, since sediment leaving most cells will eventually reach and be transported by the river to the outlet. However, the river plays the role of a transport channel, that is, it does not really influence the sediment yield attributed to a given cell, or the sediment yield reduction produced when that cell is afforested. This means that all sediment that leaves a cell and reaches the river will be fully transported to the outlet of the catchment, at least for the parameter values used during the tests, especially regarding retention capacity and saturation threshold. It is expected that using larger solution sizes would lead to an increased probability of spatial interaction occurrence among selected cells. In that case, it can be foreseen that on-site CAMF would produce significantly different results with respect to original CAMF, also this claim is not backed up by the output of on-site CAMF for Tabacay500.

Regarding execution times of on-site CAMF, it is clear that they are not influenced by solution size, since once the ranking is built it takes about the same time to select any number of cells from it. On the other hand, execution times are indeed influenced by the database size, since the time spent building the ranking of cells will depend on the number of candidate cells. CPU time fractions show the dramatic reduction on execution time that is observed when spatial interaction is left out of consideration and when all cells are selected in a single step, instead of using iterative selection.

IV. Conclusions

[6] proposed a technique called CAMF with the aim of selecting from a rasterized database representing a river catchment a set of cells to be afforested in order to minimize the sediment yield of the whole catchment. In this paper an implementation of CAMF was produced and its performance was tested on three databases representing nested river catchments in the southern Andes of Ecuador, with the aim of analyzing the behavior of CAMF when applied to databases that differ greatly in size. In addition, the influence of the number of cells to be selected on the performance of CAMF was assessed.

In contradiction to what was initially expected, the number of cells selected at each iteration by original CAMF was not exactly 1 in all tests. This indicates that the possibility of two or more cells having exactly the same sediment yield reduction value at a given iteration, although limited, does exist. However, since the observed deviation from 1 is small or, in other cases, the number of selected cells per iteration is exactly 1, execution times increase almost in direct proportion with respect to solution sizes. Besides solution size, the number of cells comprised in the database has also a clear impact on execution times. This fact allows to conclude that execution time can become a limiting factor for original CAMF, specifically in cases in which it is applied to high resolution databases covering large extents and using large solution sizes. This restriction would be even more apparent in such contexts when several runs of original CAMF are necessary, as it could be the case when performing scenario analysis, or when using original CAMF as a component of an integral model or method that requires to execute it repeatedly in a systematic way.

A variant of CAMF called on-site CAMF was also proposed, implemented and tested on the same databases as the original CAMF. On-site CAMF uses only local cell information, i.e., sediment reduction and slope, to score and rank cells. Tests using on-site CAMF produced very similar results with respect to original CAMF outputs, in an almost negligible, constant execution time. One interpretation of this finding may be that for these specific combinations of databases, solution sizes, and parameter values, spatial interaction does not play a role. This observation can be attributed to the fact that solution sizes used in tests are limited when compared to the full database sizes. It is assumed then that, when larger solution sizes are used, the relevance of the spatial interaction role will significantly increase. It is expected that, in such cases, on-site CAMF would produce different results with respect to original CAMF.

TABLE III. Output of tuning procedure for relative importance values for on-site CAMF

Solution size	$w_f = 0.5$	$w_f = 0.75$	$w_f = 0.25$	$w_f = 0.1$	$w_f = 0.01$
	$w_e = 0.5$	$w_e = 0.25$	$w_e = 0.75$	$w_e = 0.9$	$w_e = 0.99$
Tabacay500					
1	0.91	0.90	0.91	0.91	1.00
10	0.95	0.94	0.96	0.98	1.00
100	0.89	0.81	0.99	0.99	0.99
Tabacay					
1	1.00	1.00	1.00	1.00	1.00
10	0.80	0.57	1.00	1.00	1.00
100	0.60	0.42	0.88	0.99	0.99
1000	0.70	0.57	0.88	0.96	0.99
Paute					
1	1.00	1.00	1.00	1.00	1.00
10	1.00	0.99	1.00	1.00	1.00
100	0.99	0.98	1.00	1.00	1.00
1000	0.98	0.93	0.99	1.00	1.00

TABLE IV. Performance measures corresponding to on-site CAMF

Solution size	SYR $[ton\, yr^{-1}]$	SYR fraction	CPU time [s]	CPU time fraction	Spatial coincidence
Tabacay500 (initial SY: 370 $ton\, yr^{-1}$, total cells 1892, candidate cells 500)					
1	0.498	1.00	<0.001	0.000	1.00
10	4.971	1.00	<0.001	0.000	1.00
100	46.398	0.99	0.015	0.138	0.99
Tabacay (initial SY: 29075 $ton\, yr^{-1}$, total cells 68123, candidate cells 26850)					
1	3.308	1.00	0.062	0.265	1.00
10	32.171	1.00	0.062	0.033	1.00
100	197.504	0.99	0.046	0.003	0.98
1000	913.868	0.99	0.093	0.001	0.97
Paute (initial SY: 3212203 $ton\, yr^{-1}$, total cells 5616679, candidate cells 1647304)					
1	14.729	1.00	3.135	0.023	1.00
10	147.205	1.00	3.088	0.002	1.00
100	1470.557	1.00	3.634	0.000	0.98
1000	14675.398	1.00	5.834	0.000	0.97

It is clear that CAMF behavior depends heavily on the values set for its parameters. This is especially true for the retention capacities and saturation thresholds for every cell. Values for these and other parameters must be carefully determined, in order for CAMF to reproduce real world phenomena in a valid way. A systematic and scientific sound calibration procedure becomes a requirement in this regard. However, such a procedure most likely would involve a more detailed consideration of sediment production and transport, which lies beyond the scope of this paper.

References

[1] A. Molina, G. Govers, J. Poesen, H. Van Hemelryck, B. De Bièvre, and V. Vanacker, “Environmental factors controlling spatial variation in sediment yield in a central andean mountain area,” Geomorphology, vol. 98, no. 3-4, Jun. 2008, pp. 176–186.

[2] A. Palmieri, F. Shah, and A. Dinar, “Economics of reservoir sedimentation and sustainable management of dams,” Journal of Environmental Management, vol. 61, no. 2, 2001, pp. 149 – 163.

[3] FAO, “A provisional methodology for soil degradation assessment,” 1979.

[4] R. P. C. Morgan, Soil Erosion and Conservation. Blackwell Publishing, 1988.

[5] J. Yoo, S. Simonit, J. P. Connors, A. P. Kinzig, and C. Perrings, “The valuation of off-site ecosystem service flows: Deforestation, erosion and the amenity value of lakes in prescott, arizona,” Ecological Economics, vol. 97, 2014, pp. 74 – 83.

[6] P. Vanegas, D. Cattrysse, and J. Van Orshoven, “Allocating reforestation areas for sediment flow minimization: an integer programming formulation and a heuristic solution method,” Optimization and Engineering, vol. 13, 2012, pp. 247–269.

[7] J. F. O’Callaghan and D. M. Mark, “The extraction of drainage networks from digital elevation data,” Computer Vision, Graphics, and Image Processing, vol. 28, no. 3, Dec. 1984, pp. 323–344.

Service Cascade as a Means for Pan-European Access to National Geodata Content

CASE: European Location Framework

Lassi Lehto, Pekka Latvala and Jaakko Kähkönen
Department of Geoinformatics and Cartography
Finnish Geodetic Institute
Masala, Finland
lassi.lehto@fgi.fi, pekka.latvala@fgi.fi, jaakko.kahkonen@fgi.fi

***Abstract*—Providing European-wide access to geospatial data resources held on national level is the ambitious goal of the INSPIRE process and other similar integration initiatives. Service cascade is presented in this paper as a solution for facilitating access to national content to support Pan-European applications. The paper presents as an example the process of creating a map tile cache of geospatial data requested from INSPIRE-compliant Download Services. Cascading service architecture is found to be a useful approach in the cache seeding process. The work has been conducted in the context of a major EU-project.**

Keywords-service cascade; content integration; map tile cache; cross-border applications

I. INTRODUCTION

According to the fundamental INSPIRE (Infrastructure for Spatial Information in the European Community) principles, Pan-European geodata-related services are supposed to be built on top of national Member State services [1]. The same basic principle holds in the national context, where state level offerings are in many cases dependent on the availability of content and services on local government level. The traditional solution for this issue is to collect content from the local sources into a centrally managed data store and provide higher level services from that aggregated resource. In the rapidly changing world, this approach has certain drawbacks. One of the most serious problems is that the aggregated data sets fast become outdated. Various methods have been developed for incremental update of service databases from the original sources, but satisfactory results are yet to be achieved. Recently, Open Geospatial Consortium (OGC) has initiated a work related to the topic with the title GeoSynchronization, but software implementations of this work are still largely missing [2].

The concept of service cascade is evaluated in this paper as the solution for the data aggregation problem. The basic idea in service cascade is that a service can be configured as a content source for another service, actually making this latter service a client of the first service. Service cascade can be regarded as the most fitting implementation of the basic principle, according to which higher level Spatial Data Infrastructures (SDIs) can be built on top and making use of lower level SDIs. This can also be seen as the most cost-effective way for building services on multiple administrative hierarchy levels. Previously service cascade has been studied for instance in the context of metadata services integration [3].

The research described in this paper has been conducted in the context of a major EU project, called European Location Framework (ELF), initiated by EuroGeographics (EG), the co-operation organization of the European National Mapping and Cadastral Agencies (NMCAs) [4]. The ELF project aims at developing European-wide INSPIRE-compliant services from the geodata resources maintained by the EG's membership.

The ELF project started in March 2013 and will run for three years. The project has 30 participant organizations and 13 of them represent EU/EFTA member states as official NMCAs. Thus, the project has quite extensive spatial coverage extending from Finland to Spain and from Great Britain to Poland.

The ELF project includes a work package specifically dedicated for data provision and service development. In this work package there is a subtask responsible for investigating the issues related to service cascade. The approach presented in this paper covers the first phase of this development and focuses specifically on the provision of European level view services based on data services delivering pre-aggregated content.

In Section II the origins of the service cascade approach in standardization are described. In Section III the concept of map tile cache is discussed, together with the related standardization work. Section IV describes the phase one cascading service architecture adopted in the ELF project. Section V details the contents of the tile cache resulted from the service cascade work in the project and Section VI explains the challenges faced and solutions found in this process. Section VII concludes the paper, together with a short discussion about plans on the further development of the service cascade approach in the ELF project.

II. SERVICE CASCADE

The idea of service cascade has been present in the OGC's service specifications since the early days. Already the Web Map Service (WMS) version 1.0.0 introduced the functionality to define another WMS as the content source for a certain layer of information that a given WMS serves

[5]. In addition to just combining the content offerings of other services with its local map layers, the cascading WMS can provide other functionalities like coordinate system transformations, new image formats, or even more advanced image manipulation or analysis operations. Service cascading can also be used to make use of the WMS standard versions that are not supported by the original back-end service.

In the Styled Layer Descriptor (SLD) specification, the service cascading approach has been further elaborated by introducing a concept of Component WMS, a service instance that can be freely associated with any back-end content source [6]. Typically, these kinds of service combinations represent the case, in which a Web Feature Service (WFS) source is accessed by a WMS instance to provide a visual representation of the geospatial data content available in the background service.

III. Map Tile Cache

A. General

The concept of map tile cache has been introduced as the means for improving the performance of map delivery to the Web-based client applications. The idea is to pre-render at least the most frequently requested parts of the back-end data store in a form of a regularly gridded map tile cache. The most typical map tile image size is 256 * 256 pixels.

When a client application sends a request for a certain area, the tile service first checks, whether those map tiles already exist in the cache or not. If they are already in the cache, the tiles are returned right away to the calling application. If the tiles have not yet been rendered, the content for these tiles are requested from the back-end service, rendered, stored to the cache and returned to the client. This way the cache will automatically get populated for the parts that are actually needed and requested by clients. On the other hand, map tile cache implementations typically provide functionalities for fully pre-filling the cache from the data store. This process is called the feeding of the cache.

B. Tile Cache standards

Various standards have been developed as the specification for the map tile cache access interface. These include WMS Tile Caching (WMS-C) of the Open Source Geospatial Foundation (OSGeo), and the most recent Tile Map Service Specification (TMS) of the OSGeo foundation [7]. The most official one is the Web Map Tile Service (WMTS) 1.0.0 standard of the OGC [8]. In addition, the map tile service might also support the traditional WMS interface, but this requires that the service is able to assemble a single image, corresponding to an arbitrary query window, from the stored tiles.

Another important standard related to a map tile cache is the grid used to delimit the individual tiles. This grid must be bound to a Coordinate Reference System (CRS) and declare the origin and the extent of the area covered. A complete tile cache specification also includes information of the used tile size in the ground scale. Because the cache typically supports various different display scales, the contents of the cache will be generated on several zoom levels, effectively making the cache a hierarchical, pyramid-like structure. The tile size must be defined for each of the zoom levels. Often this is expressed as list of resolutions, i.e., the grid cell sizes in ground units.

IV. Service Architecture

In the first phase, the cascaded services approach in the ELF project is realized as a map tile cache containing theme-wise visualizations of the pre-aggregated Pan-European data sets Euro Global Map (EGM), Euro Regional Map (ERM) and Euro Boundary Map (EBM). Thus, in this stage the cascading service approach is specifically adopted in the tile cache seeding process. The cascaded service architecture consists of several components (Figure 1). Data content is provided by a set of WFS service instances maintained by the German NMCA, BKG, in Leipzig. The data sets have been pre-transformed to the INSPIRE-compliant form and are thus available according to the themes and feature classes, as defined in the INSPIRE data specifications.

The tile cache is implemented using Open Source software products and deployed as an Amazon Web Services (AWS) -based cloud service. The service consists of the following main components: 1) a WMTS-compliant service implementation (based on a product called MapCache), 2) a WMS service instance (MapServer), responsible for the map rendering process, 3) a seeding script, customized in the project and utilizing the ‘`mapcache_seed`’ process.

Because the used map rendering engine only supports WFS version 1.0.0 and simple features data model as the back-end source, a custom-built middleware module had to be introduced. Over time this module has been further developed to also support other functionalities. The middleware component runs on the service platform of the Finnish Geodetic Institute (FGI) in Finland.

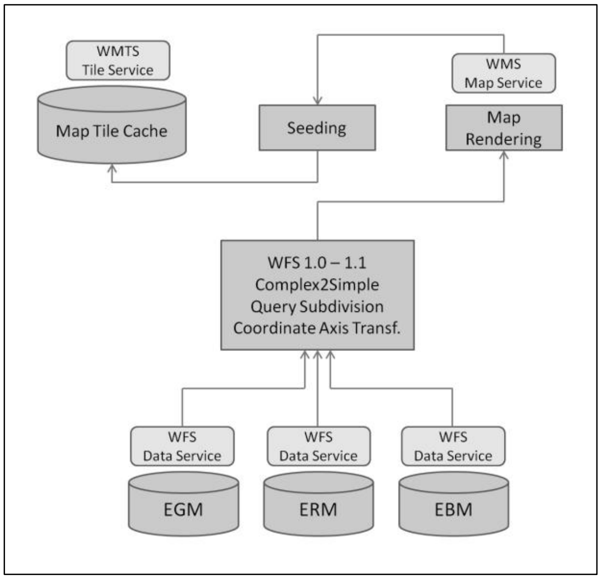


Figure 1. Service architecture of the ELF cascading services, phase 1.

Functionality of the middleware module covers four main tasks. Firstly, it takes care of the differences between the WFS version 1.0.0, supported by the map rendering engine as the source data source, and the WFS version 1.1.0 available in the back-end services. Secondly, it transforms the complex INSPIRE schemas into a simpler form that can be digested by the rendering engine. Thirdly, it performs a recursive query window subdivision, needed to resolve the timeout problem of the back-end services in case of very large polygonal features. Finally, the middleware component flips the coordinate axis order of the incoming data sets. This is required, as the treatment of CRS-related details is different in WFS 1.0.0, when compared with that in WFS 1.1.0.

V. Contents of the Tile Cache

At the moment the ELF map tile cache contains six INSPIRE-compatible map layers listed in Table I.

TABLE I. ELF Map Tile Cache Contents

INSPIRE Map Layer
AU.AdministrativeBoundaries
HY.PhysicalWaters.Watercourse
HY.PhysicalWaters.StandingWater
HY.PhysicalWaters.LandWaterBoundary
TN.RoadTransportNetworks.RoadLink
TN.RailTransportNetwork.RailwayLink

All these map layers are cached in five zoom levels that correspond to the grid sizes standardized in the INSPIRE process (100 km and 10 km), enhanced with some additional levels to facilitate smooth zooming in map viewing applications. The five zoom levels, the corresponding grid sizes, resolutions and map display scales are shown in Table II.

TABLE II. ELF Map Tile Cache Zoom Levels

Level	Grid size [km]	Resolution [m]	Appr. Scale
0	200 * 200	781.25	1 : 3 M
1	100 * 100	390.625	1: 1.4 M
2	40 * 40	156.25	1 : 500 000
3	20 * 20	78.125	1 : 280 000
4	10 * 10	39.0625	1 : 140 000

An exemplary display of the current map tile cache contents is shown in Figure 2. This includes all currently available map layers in the area of Southern Portugal. The map shown is based on Euro Global Map content in approximate scale of 1 : 3 M.

The available map layers are displayed using the INSPIRE-defined simple default styles. All layers are rendered with transparent background. This is to ease overlaying of individual map layers on top of various different types of geospatial content by third-party applications. Place names are missing from the current content offering of the ELF map tile cache, but can be integrated form the EuroGeoNames service, if needed [9].

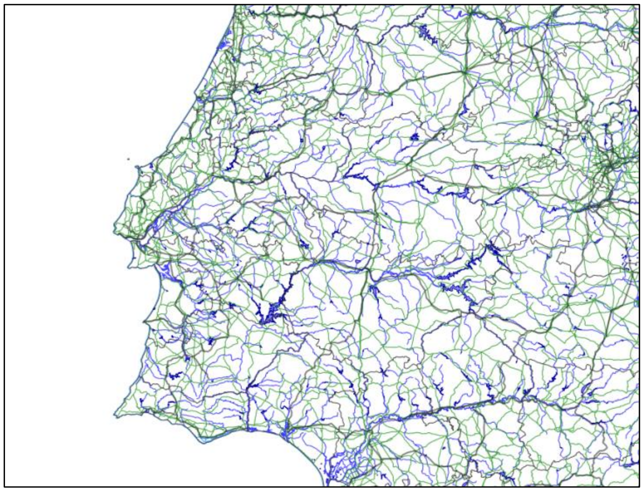


Figure 2. An exemplary map display containing all the available map layers of the ELF map tile cache, from Southern Portugal

The content in the ELF map tile cache is rendered in the Lambert Equal Area CRS (EPSG:3035) and covers the whole continental Europe, having the spatial extent of N_{min}:1300000, E_{min}:2600000, N_{max}:5500000, E_{max}:6000000.

VI. Performance and Feasibility Considerations

User-experienced performance is one of the main motivations for the development of map tile cache-based service architectures. Map tile creation is often run as a pre-processing step, called seeding of the cache. This process can be partial, i.e., carried out only in areas having most demand for map visualizations. In this approach, the areas requested less frequently will be accessed from the original source and rendered on-the-fly. User-experienced performance is naturally poorer in those areas.

In the case of the ELF project, a complete pre-seeding of the tile cache is performed. Full seeding enables the use of very poor performance data sources while keeping the user-experienced responsiveness of the system well inside the limits established in the INSPIRE regulation on service performance. In the ELF project the approach is to make use of WFS-based data download sources for map rendering process. Even complicated feature geometries can be successfully processed in this approach.

However, in the case of the most voluminous geometries, additional facilities had to be developed. In the ELF project the main reason for this was the fact that the source data services would time out on queries that involve several very complicated polygonal features, like those present in the ‘AdministrativeUnit’ and ‘StandingWater’ feature classes. Even most careful adjustment of the corresponding timeout thresholds in the middleware service and the rendering engine would not resolve this issue. Special functionality was developed in the middleware service component to tackle the problem.

When encountering the back-end service time out situation, the middleware component divides the query window into four sub-windows and re-initiates the data query as four sub-queries. If the four requests are successful, the middleware component integrates the corresponding responses as a single result data set that is returned to the rendering process. If any of the sub-queries fails due to the time out of the back-end service, the subdivision function is called recursively until a successful result is achieved. Duplicate features created in this process are ignored as they do not harm the rendering process.

VII. Conclusion and outlook

The map tile cache approach used in the ELF project enables good end user experience of the map application performance, even though low performance data source services are used. The approach is an implementation of the service cascade architecture, in which a service acts as a client to other services. The created map tile cache serves content according to the principles established in the INSPIRE process. Thus, content is divided into individual map layers and made available in the CRS and grid, as defined in the INSPIRE process. Map tiles have a transparent background to facilitate integration with other data sources.

In future, the cascading services subtask of the ELF project will focus on providing cascading functionality for the data download operations. The goal is to support the end user in accessing data content available both as pre-aggregated on European level, and directly from national services. Thus, the cascading approach aims at supporting real-time aggregation of content from a set of distributed national data sources.

One of the new challenges encountered when accessing national services from European-level applications is the need to introduce spatial integration capabilities to the traditional service cascade approach. At the moment only thematic integration is supported in the existing cascade mechanisms of the OGC service implementations. In this setup every single map layer is designated to be served by one and only one back-end service. When implementing cascaded integration over a set of national services, one has to resolve the problem of spatial query distribution and cross-border fusion of map images. These are the biggest challenges in the further service cascade developments of the ELF project, too.

Cascading of data level services also raises new challenges, because some European NMCAs do not allow caching of data content outside their organizational firewalls. In addition, many NMCAs provide access to data download only in file-based forms.

One of the preliminary plans concerning the cascading of file-based data services is to provide access to content through an interactive map user interface. According to INSPIRE, file download services are supposed to be available as Atom feeds. From the ‘coverage’ metadata element in the Atom feed, one can retrieve the spatial extent of the data in the file and display it as a rectangle on the map. This would facilitate the downloading task considerably.

A more advanced approach would be to facilitate access to file-based download services through the individual tiles of the map tile service. This would require overlap of the tile extent with available file content extents and on-the-fly selection of the portion of each file that is inside the map tile. Functionalities for this have not yet been designed or developed in the ELF project.

Another possible functionality for a cascading service in the context of data download is to convert the INSPIRE-compatible Geography Markup Language (GML) -encoded content coming from the back-end download services into a more Web-friendly format, like JavaScript Object Notation (JSON). This reformatted content could then be made available via some modern Web application-oriented access interface.

Acknowledgment

The ELF project has been funded by the European Commission, Grant Agreement No 325140.

References

[1] European Commission, INSPIRE Directive, at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108 :0001:0014:EN:PDF [accessed: 2014-01-22]

[2] Open Geospatial Consortium, Specifications Working Group on GeoSynchronization Home Page, at: http://www.opengeospatial.org/projects/groups/geosyncswg [accessed: 2014-01-22]

[3] Y. Deng and Q. Wu, “Research on the harvest and cascade of catalogue service in GeoGlobe Service Platform,” Proc. 18th International Conference on Geoinformatics: GIScience in Change (Geoinformatics 2010) Peking University, Beijing, China, June, 18-20, 2010.

[4] European Location Framework (ELF), Project Home Page at: http://www.elfproject.eu [accessed 2014-01-22]

[5] Open Geospatial Consortium, “OpenGIS Web Map Server Interface Implementation Specification, Revision 1.0.0”, April 2010, at: http://portal.opengeospatial.org /files/?artifact_id=7196 [accessed: 2014-01-22]

[6] Open Geospatial Consortium, “Styled Layer Descriptor profile of the Web Map Service Implementation Specification, Version 1.1.0”, June 2007, at: http://portal.opengeospatial.org/files/?artifact_id=22364 [accessed: 2014-01-22]

[7] Open Source Geospatial Foundation, Home page at: http//www.osgeo.org [accessed: 2014-01-22]

[8] Open Geospatial Consortium, “OpenGIS Web Map Tile Service Implementation Standard, Version 1.0.0”, at: http://portal.opengeospatial.org/files/?artifact_id=35326 [accessed: 2014-01-22]

[9] L. Lehto, P. Latvala and J. Kähkönen, “An Implementation of the OGC’s WFS Gazetteer Service Application Profile, CASE: The EuroGeoNames Central Service Renewal,” Proc. The Fifth International Conference on Advanced Geographic Information Systems, Applications and Services, ”GEOProcessing 2013”, Feb 24 – Mar 1, 2013, Nice, France.

Privacy Concerns in Location-based Social Networks

Fatma S. Alrayes, Alia I. Abdelmoty
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
{F.S.Alrayes, A.I.Abdelmoty}@cs.cardiff.ac.uk

***Abstract*—User location data collected on Location-Based Social Networks (LBSN) can be used to enhance the services provided by those applications. However, it can be potentially utilised for undesirable purposes that can compromise users' privacy. This paper presents a study of the privacy implication of location-based information provision and collection in LBSN. The study is supported by analysis of representative data sets from such applications. The results demonstrate the need for further work on improving the visibility of the information collected to users of the Social Web, to allow them to better assess the implications of their location sharing activities.**

Keywords-location privacy; LBSN; location-based inference

I. INTRODUCTION

Massive interest in geographical referencing of personal resources is evident on the Web today. Geographic referencing has evolved to become a natural method of organising and linking information with the aim of facilitating its discovery and use. GPS-enabled devices are enabling individuals to store their mobility tracks, tag photos and events. Embracing these new location-aware capabilities by the social networks has led to the emergence of Geo-Social Networks (GeoSNs) which offer their users the ability to geo-reference their submissions and to share their location with many other users. Subsequently, users can use the location identifier to browse and search for resources. GeoSNs include Location-Enabled Social Networks (LESNs), for example, Facebook, Twitter and Flickr, where users' location is supplementary identification of other primary data sets, and Location-Based Social Networks (LBSNs), for example, Foursquare and Yelp where location is an essential key for providing the service,

GeoSNs collect real-time and large-scale location information on users as well as other contextual information including user relationships and user provided text updates possibly over long periods of time. In particular, LBSNs as opposed to LESNs enable sharing and collection of detailed personal location information, provide significant semantic data associated with the location, such as place name, type and address, as well as allow users to express their opinions and experience in terms of reviews and tips. As a result, users' historical location information can be related to contextual and semantic information publicly available online [1] and can be used to infer personal and sensitive information about users and for constructing comprehensive user profiles. Possible derived information in such profiles can include user activities, relationships, interests and mobility patterns [2][3]. Such enriched location-based profiles can be considered to be useful if used to personalise and enhance the quality of use of the applications. However, they can potentially be used for undesirable purposes and can pose privacy threats ranging from location-based spams to possible physical harm by any adversary [4]. Users may not be fully aware of what location information is being collected, how the information is used and by whom, and hence can fail to appreciate the possible potential risks of disclosing their location information.

In this paper, we study the location privacy of users when using LBSNs. The aim of this study is to investigate the potential privacy implications of LBSNs by examining and demonstrating possible derived information from typical data sets collected by these applications for different types of users. Foursquare was chosen as a representative LBSN application for conducting this study due to its popularity (over 40 million users across the world by September, 2013 [5]).

Firstly, the dimensions of the location privacy problem in LBSNs are examined in terms of the type of data collected, its visibility and accessibility by users of the application, as well as the possible exploitation of these data and the level of security in such services, in order to provide a better understanding of the privacy issues. Secondly, an analytical study is carried out, using a representative data set, to explore the location data content and the range of possible inference that can be made from them. Usage patterns in the dataset are used to guide a classification of users of the application and in the analysis of the data.

Previous studies focused mostly on examining spatiotemporal movement patterns in LBSNs [6][7]. Some studies explored users' privacy concerns and attitude when sharing their location for social purposes, but presented limited evaluations using restricted application scenarios [8][9].

This paper presents a study of the privacy implications of location-based information provision and collection in LBSN. The study is supported by analysis of representative data sets from such applications. The results demonstrate the need for further work on improving the visibility of the information collected to users of the Social Web, to allow them to better assess the implications of their location sharing activities.

The rest of this work is organized as follows: section II provides an overview of related work. Section III discusses the dimensions of the location privacy problem in LBSNs. Section IV describes the experiment conducted. Section V presents the result of the analyses. A conclusion of the work and future directions are given in Section VI.

II. Related Work

Two relevant questions to the problem being studied are: to what extent is location privacy a potential concern for users in LBSNs, and what sort of location-based inference is possible from the data collected by LBSNs. In this section, related work on both issues is reviewed.

A. Users Attitudes and Privacy Concerns in Geo-Social Applications

A growing research interest has been witnessed over the past few years for studying users' attitudes and privacy concerns of their location privacy and investigating how user-empowered location privacy protection mechanism can influence their behaviour. Tsai et al. [8] developed a social location sharing application where participants were capable of specifying time-based rules to share their location and they were then notified of who viewed their locations. Their findings suggested that the control given to user for setting their sharing preferences has contributed to the reduction of the level of privacy concern of the participants.

Similarly, Sadeh et al. [9] enabled users of their People Finder application to set rule-based location privacy controls by determining the where, when and with whom to share their location and were notified when their location information was requested. Participants were initially reluctant to share their location information and then tended to be more comfortable over time. Patil et al. [10] implemented a system that represents actual users' workplace offering live feeds about users and their location, then asked those users to define permissions for their personal information sharing by setting various levels linked to four user categories. They found that these participants were concerned most about their location information and they utilised the permission feature to control it. Another study by Kelley et al. [11] showed that users were highly concerned about their privacy especially when sharing location with corporate-oriented parties. However, their location-sharing with advertising companies can be increased when offering more complex location privacy settings.

Other work was carried out to examine how the employment of visualisation methods may impact users' attitude to location privacy and behaviour. Brush et al. [12] studied users' attitudes towards their location privacy when socially sharing their location or when tracking it using GPS for long periods of time and questioned whether using some obfuscation techniques can address users' concerns. As a result, participants were concerned about revealing their home, identity and exact locations. They visually recognised and chose the best obfuscation techniques they felt protect their location privacy. In addition, Tang et al. [13] investigated to what extent presenting various visualisations of users' location history can influence their privacy concerns when using location sharing applications. They developed text-, map-, and time-based visualisation methods and considered spatiotemporal properties of sharing historical location. They noted that the majority of participants were concerned about their location privacy including their physical privacy when showing them their visualised location history and consequently preferred text-based visualisation when sharing location with other users, as it was perceived to limit the amount of information exposed.

With regards to public GeoSNs, there is relatively few research works that examines privacy concerns of users. Lindqvist et al. [14] considered users' motivations in using Foursquare and questioned their privacy concerns. Their analysis showed that most of the participants had few concerns about their privacy and users who were more concerned about their privacy chose not to check into their private residence or to delay checking into places till after they leave, as a way of controlling their safety and privacy. A similar observation was noted by Jin et al. in [15], where it was found that users were generally aware of the privacy of their place of residence and tended not to provide full home addresses or blocked access to their residential check-ins to other users.

In summary, it is evident that location privacy presents a real concern to users in location sharing applications, and particularly as they become aware of the data they are providing. Previous studies may have been limited by several factors, including the size and representativeness of the sample user base used in the experiments conducted and the limited features of the proprietary applications used in testing [8][9][10][11]. Moreover, as far as we are aware, there are no previous studies that consider the problem of location privacy on public LBSNs.

B. Location-Based Inference from GeoSNs

There are some studies that utilised publicly available information from GeoSNs in order to derive or predict users' location. In [16], Twitter users' city-level locations were estimated by only exploiting their tweets contents with which it was possible to predict more than half of the sample within 100 miles of their actual place. Similarly, Pontes at al [17] examined how much personal information can be inferred from the publicly available information of Foursquare users and found the home cities of more than two-third of the sample within 50 kilometres. Sadilek at al. [18] investigated novel approaches for inferring users' location at a given time by taking advantage of knowing the GPS positions of their friends on Twitter. Up to 84% of users' exact dynamic locations were derived. Interestingly, Gao et al. [19] formulated predictive probability of the next check-in location by exploiting social-historical ties of some Foursquare users. They were able to predict with high accuracy possible new check-ins for places that users have not visited before, by exploiting the correlation between the social network information and geographical distance in LBSNs [20].

Other works focussed on investigating the potential inference of social relationships between users of GeoSNs. Crandall et al. [21] investigated how social ties between people can be derived from spatial and temporal co-occurrence by using publicly available data of geo-tagged pictures from Flickr. They found that relatively limited co-occurrence between users is sufficient for inferring high probability of social ties. Sadilek at al. [18] also formulated friendship predictions that derive social relationships by considering friendship formation patterns, messages' content of users and their location. They predicted 90% of friendships with accuracy beyond 80%. Additionally, Scellato et al. [22] investigated the spatial properties of social networks existing among users of three popular LBSNs and found that the likelihood of having social connection decrease with distance. Scellato et al. [23] developed a link prediction system for LBSNs by utilising users' check-ins information and places properties. 43% of the all new links appeared between users that have at least one check-in place in common and especially those who have a friend in common.

Studying and extracting spatiotemporal movements and activities patterns of users on GeoSNs have also attracted much research in recent years. Dearman et al. [24] exploited locations' reviews of Yelp in order to identify a collection of potential activities promoted by the reviewed location. They derived the activities supported by each location by processing the review text and validated their findings through a user questionnaire. Noulas at al. [25] studied user mobility patterns in Foursquare by studying popular places and transitions between place categories. Cheng at al. [6] examined a large scale dataset of users and their check-ins to analyse human movement patterns in terms of spatiotemporal, social and textual information associated with this data. They were able to measure user displacement of consecutive check-ins, distance between users' check-ins and their centre of mass, and the returning probability to venues. They also studied the factors affecting users' movement and found considerable relationship between users' mobility and geographic and economic conditions. More recently, Preotiuc-Pietro et al. [7] investigated the behaviour of thousands of frequent Foursquare users. They analysed users' movements including returning probability, check-ins frequency, inter-event time, and place transition among each venue category. They were also able to group users based on their check-in behaviour such as generic, businessmen or workaholics as well as predicting users' future movement.

The above studies show that there is a significant potential for deriving personal information form GeoSNs and hence imply the possible privacy threats to user of these applications. Whereas previous studies considered mobility and behaviour of large user groups and determined general patterns and collective behaviour, in this work we consider the privacy implications for individual users, with the aim of understanding possible implied user profiles from location data stored in LBSNs.

III. Location Privacy on LBSNs

Four aspects of location privacy on LBSNs can be identified. These are related to the amount of data collected and its quality, its visibility and accessibility, its possible utilisation by potential users, and the level of security offered to the user by the application.

A. *Location Data Collection*

Location data collection refers to the type of location data collected and stored as well as to its quality.

Foursquare collects and records user location data automatically and continuously, by estimating the user's current latitude and longitude from the device being used. User's check-ins into specific places are verified against their estimated current location and recorded explicitly. Foursquare also states that it collects additional information from third parties services, that communicate with the application, including personal information and activities.

User's visit to a place is recorded by the user intentionally checking-in a place. Check-ins are made against predefined venues. Venues record detailed information about a geographic place, including a name, a type classification, coordinate point representation, and address. In addition, users' check-ins are also time stamped. Depending on the frequency of check-ins, a user (and the system) is able to record a complete and highly specific spatiotemporal track of their mobility.

B. *Location Information Accessibility*

Location information accessibility is concerned with how much of users' data are available and visible to others including the user, other users and the third parties of the service.

Users' pervious check-in information is provided to them in the form of check-in history where they can view their visited venues, date of the visit and any tips they have made. They are also able to access and download their check-in. However, users seem to have only a limited aspect of accessibility compared with what service provider can collect or exploit these data. For example, Foursquare states in their privacy policy that they record users' location on a continuous bases even without users checking into venues, but users do not have access to this data.

Almost all of the users' information is publicly available by default and can be viewed by other users. This includes profile information, tips, likes, friends list, photos, badges, mayorships, and check-ins. Users are only able to block access to their check-ins and photos by setting their view to 'private'.

As for information disclosure to third parties including Foursquare API's users, all of the publicly available users' information is accessible by third parties including private users' information such as check-ins in anonymous form that is not linked to individual users. Foursquare also indicates that they will share users' personal information with their business partners and whenever is necessary in some situations, such as enforcement of law.

C. *Location Data Exploitation*

Location information exploitation refers to how the application or third parties can utilise the data and for which purposes.

Foursquare gives itself absolute privileges over using and manipulating user information as stated in their terms of use.

> "By submitting User Submissions on the Site or otherwise through the Service, you hereby do and shall grant Foursquare a worldwide, non-exclusive, royalty-free, fully paid, sublicensable and transferable license to use, copy, edit, modify, reproduce, distribute, prepare derivative works of, display, perform, and otherwise fully exploit the User Submissions in connection with the Site, the Service and Foursquare's (and its successors and assigns') business, including without limitation for promoting and redistributing part or all of the Site (and derivative works thereof) or the Service in any media formats and through any media channels (including, without limitation, third party websites and feeds)."

From the above, it is clear that there are no commitments from the application provider as to how the data may be used or shared by the application or by other parties. Hence, by agreeing to the terms and conditions, users effectively are giving away the data and unconditional rights to use the data to the application.

D. Location Data Security

Location data security refers to the level of data protection provided by the application for securing the user's data against the risk of loss or unauthorized access.

Foursquare declares that the security of users' information is not guaranteed and any "Unauthorized entry or use, hardware or software failure, and other factors, may compromise the security of user information at any time". Without any commitment to responsibility for data security, the application provider is declaring the possible high risk of data abuse by any adversary or even by the application provider themselves.

IV. Experiment

This analysis is carried out using a real-world dataset from Foursquare for the purpose of demonstrating privacy implications of user activity on LBSNs. The effect of location data density and diversity on the possible inferences that can be made from the data is also analysed.

A. Dataset

The Foursquare dataset used in this analysis is provided by Jin et al. [15]. The dataset contains venue information and public check-ins for anonymised users around the wide area of Pittsburgh, USA from 24 February, 2012 to 22 July, 2012. It contains 60,853 local venues, 45,289 users and 127,6988 public check-ins of these users.

B. Approach and Tools Used

To study the possible impact of location data density on users' privacy, users of the dataset were first classified into groups based on their check-in frequency. A filter was initially imposed to disregard sparse user activity. Hence, users with less than five check-ins per month were removed from the dataset. The rest of the users were categorised into three groups based on their check-in frequency per day, to Moderate, Frequent and Hyper-active user groups, as shown in Table I. One representative user is selected from each group who has the nearest average check-ins per day to the average check-ins per for the whole group. Table II shows some statistics for the selected users.

The R statistical package was used for analyses and presentation of results. The SQLDF package was used for querying, linking and manipulating the data and the ggplot2 package was used for the presentation of the results of the analysis.

V. Results

Analysis of the dataset questioned the sort of implicit user-related information that can be considered to be private that may be extracted using the location data collected. The user's spatial location history can be extracted, in the form of visits to venues and the exact times of such visits. The places visited are identified and described in detail. For example, user7105 visited 'Kohl's'; a department store, located at latitude 40.51105772555344 and longitude -79.99340577016872 at 9 a.m. Monday 27/2/2012.

The basic information on venue check-ins can be analysed further and combined with other semantic information from the user profile to extract further information that can compromise user's privacy. Analysis will investigate the relationship between users and places visited, their mobility patterns and the relationships between users and other users as follows:

- **Degree of association between user and place**. Relationship with individual place instances as well as with general place types or categories will be studied. Elements of interest will include visit frequency, and possible commuting habits in terms of the association between the visit frequency of places and their location.
- **Spatiotemporal movement patterns.** Visiting patterns to individual places or to groups of places can identify regular movement patterns. In addition, a change of visit patterns can also be a significant pointer to user activity.

Table I. Statistics of users' grouping.

Group Name	**Check-ins Range in Total**	**Users Count**	**Check-ins Range per Day**	**Average Check-ins per Day**
Moderate	Between 50 and 300	4902	0.3 to 2	1.15
Frequent	Between 301 and 750	880	2 to 5	3.5
Hyper-active	Between 751 and 1303	24	5 to 8.6	6.8

Table II. Statistics of the selected users.

Factor	**Selected Users**		
	User9119	***User7105***	***User2651***
Number of total check-ins	144	511	1019
Average check-ins per day	**0.96**	**3.4**	**6.8**
Number of visited venues	21	99	101
Number of visited venues' categories	17	47	57
Number of visited venues' main categories	10	11	17
Number of friends	20	10	19

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

- **Degree of association with other users.** Relationship between users can be derived by studying their movement patterns and analysing their co-occurrence in place and time.

A. The Moderate User

The analysis results of user9119 selected from the moderate groups are as follows.

1) Degree of Association Between User and Place

Two frequently visited venues by user9119 are ‘Penn Garrison’ whose category is ‘Home’ and ‘USX Tower’ whose category is ‘Office’ representing 44% and 36% respectively of the total check-ins. Home and office are highly sensitive places, yet they represent 80% of this user’s check-ins. Other visited place types with significantly less frequency include, ‘Nightlife Spot’:0.5%, ‘Travel & Transport’: 0.27%, and ‘Shop & Service’:0.27%. User9119 is also interested in ‘Hockey’, ‘Garden Center’ and ‘Museum’ place types. As could be predicted, the location of venues visited, indicates that most of the visited venues are close to ‘Home’ and ‘Office’, whereas this user commutes further away to visit some less frequent venues such as ‘Hockey Arena’. Figure 1 shows this user’s check-in frequency for different categories of venues classified by the time of day. As can be seen from the figure, user’s association with sensitive places like home and place of work can be identified. In addition, a strong association with other place categories is also evident.

2) Spatiotemporal Movement Patterns

About 40% of this user’s total check-ins occurs at 9 am, mostly in the ‘Office’ and at 7 pm, mostly at ‘Home’. More than two-thirds of the check-ins are between 10 am and 2 pm and between 6 pm and 11 pm, which indicates that this user commutes more frequently during these hours. From the user’s weekly patterns of movement, it can be seen that 71% of the venues were visited after 6 pm. Mondays and Thursdays are when this user is most active, representing 41% of the check-ins. User9119 tend to go to ‘Nightlife spots’ more frequently during working days, whereas visits to other specific place types occur only at weekends, including, ‘Salon or Barbershop’, ‘Coffee Shop’ and ‘Garden Centre’. This user typically starts commuting earlier on working days and visits more places than on weekends. Observing the check-ins by month shows that the months of May and June are the most active in terms of the check-in frequency, comprising 60% of total check-ins, as well as diversity of category of venues visited (99% of the total visited categories of venues occurred in those months, including the emergence of new categories such as ‘Museum’, ‘Airport’ and ‘Hotel’). The user was least active in April. Figure 4 demonstrates this user's check-ins count in different categories of venues, classified by day and grouped by month.

Some changes of this user’s habits can be noticed as well, which can suggest a change of the user’s circumstances. For example, the user has not visited any Nightlife spots in March and April and has not checked-in in any place on Sundays of June and July including ‘Home’ and ‘Office’. In addition, the user has not checked in any place for a period of a week between the 21st and 28th. User9119 last check-in before this week was on the 20th of April at ‘Home’. This may indicate a possible period of time-off work in that week.

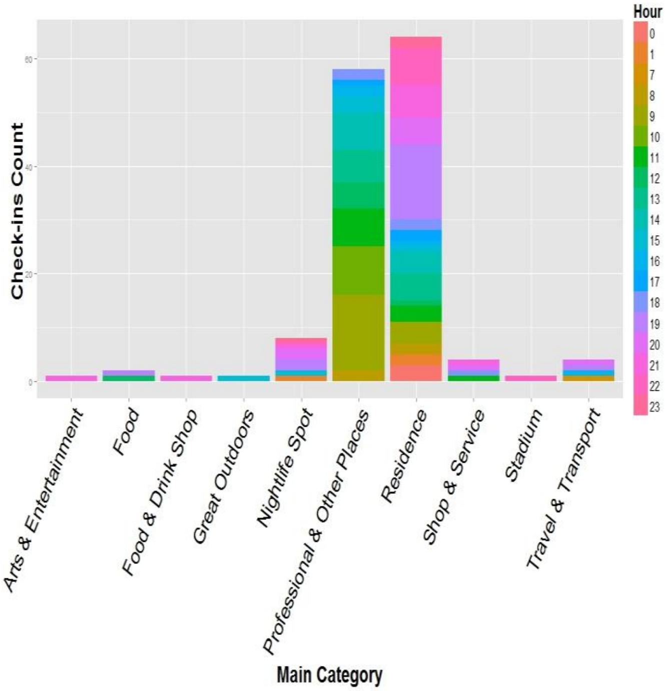


Figure 1. The moderate users’ check-ins count, classified by the category of venues for different hours of the day.

3) Degree of Association with Other Users

Co-location is used here to denote that users have visited the same venue. This can be used as a measure of uses’ interest in a place. User9119 was co-located at in 6 unique venue categories with two (out of twenty) friends.

Spatiotemporal co-occurrence between users is co-location at same place and time. This can be used as a measure of relationship between users. User9119 shared three co-occurrences with two friends; once with friend1236 at ‘American Restaurant’ and twice with friend15229 at ‘Office’, which can indicate that friend15229 is a colleague at work. In fact, this user shared 95 co-occurrences with 52 other users, 90% of which were in the ‘Office’ suggesting the probability of those users being work colleagues.

B. The Frequent User

Analysis of results of user7105 from the frequent user group is as follows.

1) Degree of Association between User and Place

Similar to the moderate user, user7105 most checked-in venue category is ‘Home’, whose location is identified in detail. However, the second most visited venue is a specific restaurant, whose category is ‘American Restaurant’, representing 25% of the total check-ins and 28% of category check-ins. This visit pattern may indicate that this is the user’s work place.

The third most visited venue category for this user is ‘Bar’ (4%), that is a subcategory of ‘Nightlife Spot’, representing about 7% of check-ins. Generally, the third most visited main category is ‘Shop & Service’ corresponding to 10% of check-ins where specifically 40% of it to ‘Gas Station or Garage’ and 25% to ‘Drugstore or Pharmacy’. User7105 occasionally interested in visiting places described as ‘Great Outdoors’, ‘Professional & Other Places’ and ‘Arts & Entertainment’.

The majority of the most frequently visited venues are within close distance to ‘Home’ and to the ‘American Restaurant’, whereas user7105 commutes further away for other less frequently visited places, such as, the ‘Medical Center’.

2) Spatiotemporal Movement Patterns

Generally, about 20% of the check-ins occurs from 10 am to 12 pm, half of which are at ‘Home’. In addition, user7105 tend to move the most between 3pm and 5pm, representing 23% of his total check-ins to 46% of the visited venues' categories. More than half of the check-ins are at ‘Atria's’, which may indicate that the user starts his work shift in this place at that time. This hypothesis can be ascertained by examining his subsequent check-ins, where 18% of the check-in happens between 12 am and 3 am at ‘Home’, possibly when the user comes back from work. There is a high correlation in terms of place transition between ‘Home’ and the ‘American Restaurant’.

When examining the weekly mobility, user7105 is more active on Tuesdays followed by Saturdays corresponding to 19% and 16% respectively of total check-ins. Noticeably, the majority of Friday and Tuesday check-ins occurs at 12 am, whereas Monday and Saturday at 4 pm. Furthermore, this user has visited more diverse venues on Tuesdays followed by Thursdays and Wednesdays representing 53%, 43% and 38% respectively of total visited categories.

During the working week, this user tend to visit a ‘Bar’ (5%), especially on Tuesdays, and ‘Gas Station or Garage’ (4%). This may be reasonable considering his working shifts. While on weekends, ‘Grocery or Supermarket’ and ‘Drugstore or Pharmacy’ venues are among the top four visited categories corresponding to 4% and 5% respectively of weekends’ check-ins.

User7105’s check-in patterns was regular over the whole period. However, this user’s visits are more frequent and diversified in the month of March. Noticeably, about 28% of the check-ins between 12 and 3 am occurred in March, indicating a possible change of lifestyle. Figure 5 presents this user's check-ins count in different categories of venues, classified by day and grouped by month.

3) Degree of Association with Other Users

User7105 had co-locations in 36 unique venues from 19 different categories with 7 friends. In particular, 26 co-locations are shared with the freind38466 at 14 venues categories including ‘Coffee Shop’, ‘Bar’, ‘Fast Food Restaurant’ and ‘Other Nightlife’. Co-locations shared with the rest of the friends include ‘Bar’, ‘Mexican Restaurant’, ‘Hospital’ and ‘Government Building’.

Moreover, user7105 has 16 spatiotemporal co-occurrences at 14 unique venues from 6 different categories with two friends where 14 co-occurrences with freind38466 at 6 different categories including mostly ‘Bar’, ‘American Restaurant’, and ‘Sandwich Place’, which can denote a close friendship between them. The other two co-occurrences are with friend15995 at ‘American Restaurant’ on May 13th and June 17th, 2012. The place and time of this user’s co-occurrences with friends are shown in Figure 2. Similarly, this user also has 89 co-occurrences with other users, who are not stated as friends, at 29 unique venues where 38% of these co-occurrences at ‘American Restaurant’ and 24% at ‘Plaza’.

C. The Hyper-Active User

The results of analysis for user2651 selected from the hyper-active user group are as follows.

1) Degree of Association Between User and Place

The first most visited venue by this user is a ‘Nightlife Spot’ corresponding to 15% of total check-ins. Two ‘Home’ venues were recorded, ‘My Back Yard’ and ‘La Couch’, representing 23% of the check-ins. Both home venues have the same location coordinates, implying that they are actually the same place. ‘Automotive Shop’, ‘Pool’ and ‘Italian Restaurant’, representing 9%, 8% and 5% respectively of this user’s total check-ins indicating the user’s interests and activities which can be swimming and Italian food. A particular instance with a vague category of ‘Building’ was among the top 10 most visited venues. Further investigation of this venue using the given place name revealed that this building is a place where an international summit for creative people is held [26]. That indicates that user2651 is possibly an active participant of such an event.

When considering the main category of the visited venues, this user generally visits ‘Shop & Service’, ‘Nightlife Spot’, ‘Arts & Entertainment’ and ‘Food’ on a regular basis, representing 17%, 14%, 11% and 10% respectively of this user’s check-ins. User2651 also usually visits ‘Gas Station or Garage’: 4%, and ‘Church’: 3%, which can imply that this user is a person with faith. The location of the visited venues can be clustered into two main areas on a map as illustrated in Figure 3. One area is where the user’s ‘Home’ is location, as well as other frequently visited venues such as ‘Nightlife Spots’ and ‘Gym or Fitness Center’. The other area includes mostly less frequently visited venues such as ‘Hospital’.

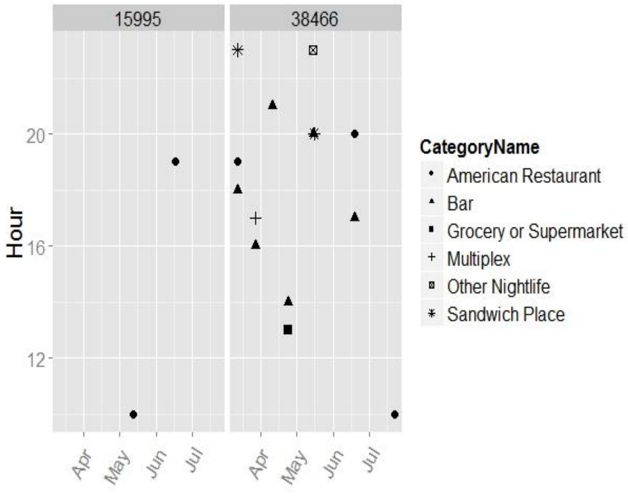


Figure 2. Spatiotemporal tracks of the frequent user co-occurrences with friends.

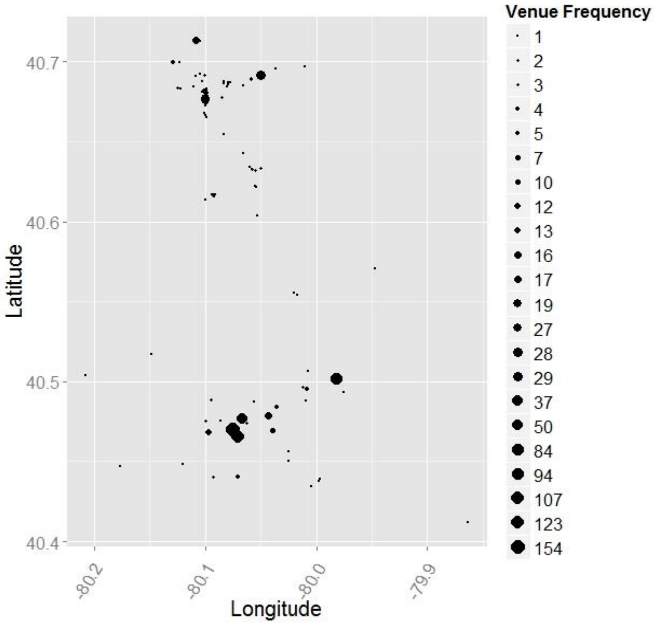


Figure 3. Venues' coordinates visited by the hyper-active user by considering the frequency of visit.

2) *Spatiotemporal Movement Patterns*

Overall, 53% of residential check-ins occurs between 9 am and 12 pm where user651. This user’s check-in frequency reaches the peak at 2 pm where 10% of the check-ins occurs and about two-third of them into the ‘Automotive Shop’. Moving towards the night, user2651’s check-in frequency reached another peak between 11 and 12 am representing 18% of the check-ins in which more than half is into ‘Nightlife Spot’ where this user may work at, and a third into ‘Home’ when potentially returning home. Noticeably, this user tends to be more active at night since about 70% of the check-ins happens after 6 pm.

Surprisingly, weekends have similar check-in frequencies as working week, and Sunday has the highest higher check-in frequency among the week days which is not the expected movement habit for average people. Moreover, user2651 checks in considerably less at the ‘Automotive Shop’ and the ‘Pool’ on Wednesday and Friday respectively. However, user2651 checks in the ‘Automotive Shop’ and the ‘Nightlife Spot’ even in weekends, which may suggest that this user has weekends work shifts. In addition, this user typically has some different priorities of visit between working week and weekend. For example, ‘Church’ is the sixth most visited venue category in weekends, whereas in working days, ‘Bar’ is the sixth most visited venue category.

User2651 has regular check-in patterns over the whole period. However, in the months of June and July, the user’s check-ins into ‘Hotel’ and ‘Pool’ significantly increase representing 75% and 60% respectively of these venues total check-ins. Moreover, other categories or venues are highly visited in certain months. For instance, 35% of total ‘Gas Station or Garage’ check-ins occurs in April, which can indicate that this user commutes more at that time, and 40% of total ‘Church’ check-ins occurs in July. Figure 6 demonstrates this user's check-ins count in different categories of venues, classified by day and grouped by month.

3) *Degree of Association with Other Users*

User2651 shares co-locations in 27 unique venues from 19 categories with 9 friends where 13 co-locations are with friend12432 and 9 with friend12046. Most of co-locations with this user’s friends are in ‘Nightlife Spots’, ‘Gas Station or Garage’, ‘Pool’, ‘Flower Shop’ and ‘Bar’. This user also has 16 co-occurrences with three friends where 4 of them with friend12046 and 3 with friend12432 at a ‘Nightlife Spots’, ‘Pool’, ‘Flower Shop’ and with just friend12432 at ‘Automotive Shop’. As with other users, user2651 co-occurred with 23 users at 12 distinct venues where half of these co-occurrences happened in ‘Bar’, ‘Automotive Shop’ and ‘Grocery or Supermarket’.

VI. CONCLOSION AND FUTURE WORK

In this paper, we investigated the privacy implication of location-based information provision and collection in LBSNs. The study is supported by analysis of a representative dataset from Foursquare. The results showed that it is highly feasible to infer rich personal information about users and their mobility. In particular, some of the possible inferences demonstrated are:

- Users’ spatiotemporal movement tracks and patterns.
- Users’ absence and presence in particular places.
- Visiting frequencies and possible degree of association with specific places or place types.
- Users’ commuting habits.
- Co-location patterns with other users and friends.

More work needs to be done to investigate the following issues:

- The relationship between the density of information and the accuracy of the inference.
- The effect of the integration of users’ data from different LBSNs.
- The relationship between the amount of information that can be analysed and the users’ perception of personal privacy.

The study also demonstrates the need for further work on improving the visibility of the information collected to users of the Social Web to allow them to better assess the implications of their location sharing activities.

ACKNOWLEDGMENT

We acknowledge the use of the Foursquare dataset provided by Lei Jin; from the School of Information Sciences, University of Pittsburgh, USA.

REFERENCES

[1] C. Ruiz Vicente, D. Freni, C. Bettini, and C. S. Jensen, “Location-Related Privacy in Geo-Social Networks,” IEEE Internet Computing, vol. 15, no. 3, pp. 20–27, May 2011.

[2] D. Riboni, L. Pareschi, and C. Bettini, “Privacy in Georeferenced Context-aware Services : A Survey,” in in Privacy in Location-Based Applications, Berlin, Heidelberg: Springer-Verlag, 2009, pp. 151–172.

[3] S. Gambs, O. Heen, and C. Potin, “A comparative privacy analysis of geosocial networks,” in SPRINGL ’11 Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, 2011, pp. 33–40.

[4] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, “Quantifying Location Privacy,” 2011 IEEE Symposium on Security and Privacy, pp. 247–262, May 2011.

[5] Foursquare, “About.” [Online]. Available: https://foursquare.com/about. [Accessed: 29-Sep-2013].

[6] Z. Cheng, J. Caverlee, K. Lee, and D. Sui, “Exploring Millions of Footprints in Location Sharing Services.,” in ICWSM, 2011, vol. 2010, no. Cholera, pp. 81–88.

[7] D. Preotiuc-Pietro and T. Cohn, “Mining User Behaviours: A Study of Check-in Patterns in Location Based Social Networks,” in Web Science, 2013.

[8] J. Y. Tsai, P. Kelley, P. Drielsma, L. F. Cranor, J. Hong, and N. Sadeh, “Who ’ s Viewed You ? The Impact of Feedback in a Mobile Location-Sharing Application,” in CHI ’09 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2009, pp. 2003–2012.

[9] N. Sadeh et al., “Understanding and capturing people’s privacy policies in a mobile social networking application,” Personal and Ubiquitous Computing, vol. 13, no. 6, pp. 401–412, Oct. 2008.

[10] S. Patil and J. Lai, “Who gets to know what when: configuring privacy permissions in an awareness application,” in Proceedings of the SIGCHI conference on human factors in computing systems (CHI 2005), 2005, pp. 101–110.

[11] P. G. Kelley, M. Benisch, L. F. Cranor, and N. Sadeh, “When are users comfortable sharing locations with advertisers?,” in Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11, 2011, pp. 2449–2452.

[12] a. J. B. Brush, J. Krumm, and J. Scott, “Exploring end user preferences for location obfuscation, location-based services, and the value of location,” in Proceedings of the 12th ACM international conference on Ubiquitous computing - Ubicomp ’10, 2010, p. 95.

[13] K. P. Tang, J. I. Hong, and D. P. Siewiorek, “Understanding how visual representations of location feeds affect end-user privacy concerns,” in Proceedings of the 13th international conference on Ubiquitous computing - UbiComp ’11, 2011, pp. 207–216.

[14] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, “I’m the mayor of my house: examining why people use foursquare-a social-driven location sharing application,” in CHI ’11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 2409–2418.

[15] L. Jin, X. Long, and J. B. D. Joshi, “Towards understanding residential privacy by analyzing users’ activities in foursquare,” in Proceedings of the 2012 ACM Workshop on Building analysis datasets and gathering experience returns for security - BADGERS ’12, 2012, pp. 25–32.

[16] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: a content-based approach to geo-locating twitter users,” in CIKM ’10 Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 759–768.

[17] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida, “We Know Where You Live : Privacy Characterization of Foursquare Behavior,” in UbiComp ’12 Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012, pp. 898–905.

[18] A. Sadilek, H. Kautz, and J. Bigham, “Finding your friends and following them to where you are,” in WSDM ’12 Proceedings of the fifth ACM international conference on Web search and data mining, 2012, pp. 723–732.

[19] H. Gao, J. Tang, and H. Liu, “Exploring Social-Historical Ties on Location-Based Social Networks.,” in Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012, pp. 114–121.

[20] H. Gao, J. Tang, and H. Liu, “gSCorr: modeling geo-social correlations for new check-ins on location-based social networks,” in Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 1582–1586.

[21] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, “Inferring social ties from geographic coincidences.,” Proceedings of the National Academy of Sciences of the United States of America, vol. 107, no. 52, pp. 22436–41, Dec. 2010.

[22] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, “Socio-Spatial Properties of Online Location-Based Social Networks.,” in ICWSM, 2011, pp. 329–336.

[23] S. Scellato and C. Mascolo, “Exploiting Place Features in Link Prediction on Location-based Social Networks Categories and Subject Descriptors,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, no. Section 3, pp. 1046–1054.

[24] D. Dearman and K. Truong, “Identifying the activities supported by locations with community-authored content,” in Proceedings of the 12th ACM international conference on Ubiquitous computing, 2010, pp. 23–32.

[25] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, “An Empirical Study of Geographic User Activity Patterns in Foursquare.,” in ICWSM, 2011, pp. 70–573.

[26] “World Domination Summit.” [Online]. Available: http://worlddominationsummit.com/faq/#primary-content. [Accessed: 05-Nov-2013].

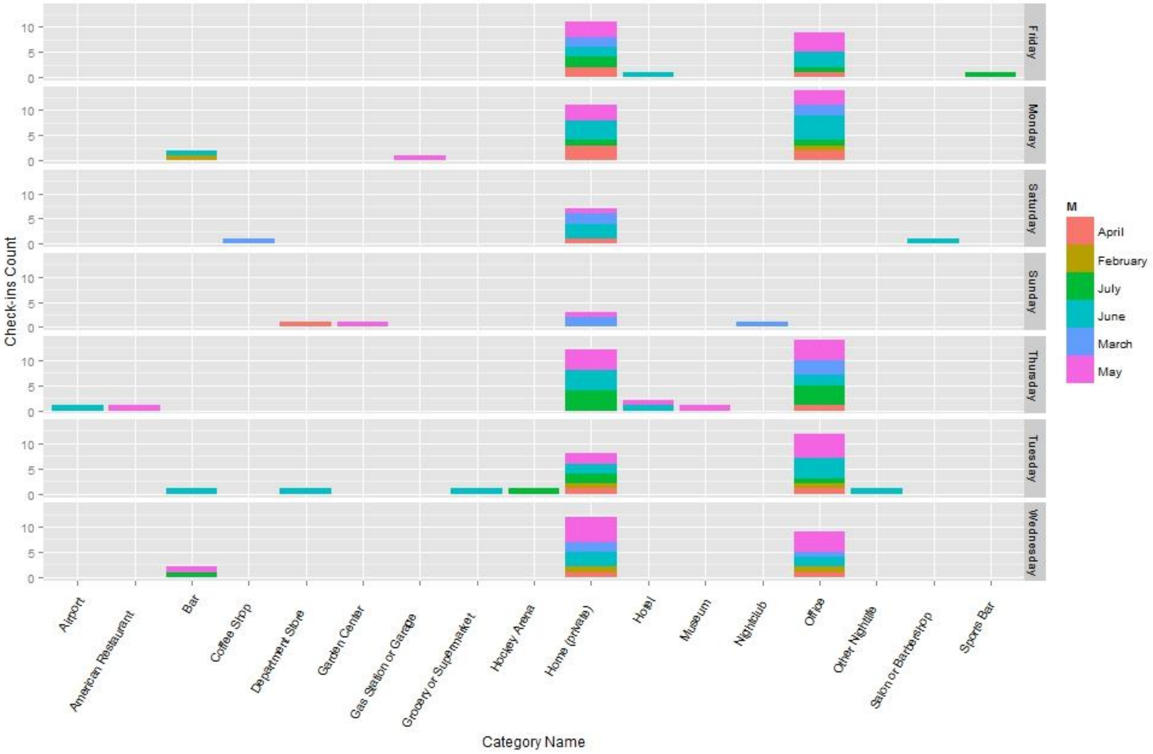


Figure 4. The moderate user's check-ins count in different categories of venues, classified by day and grouped by month.

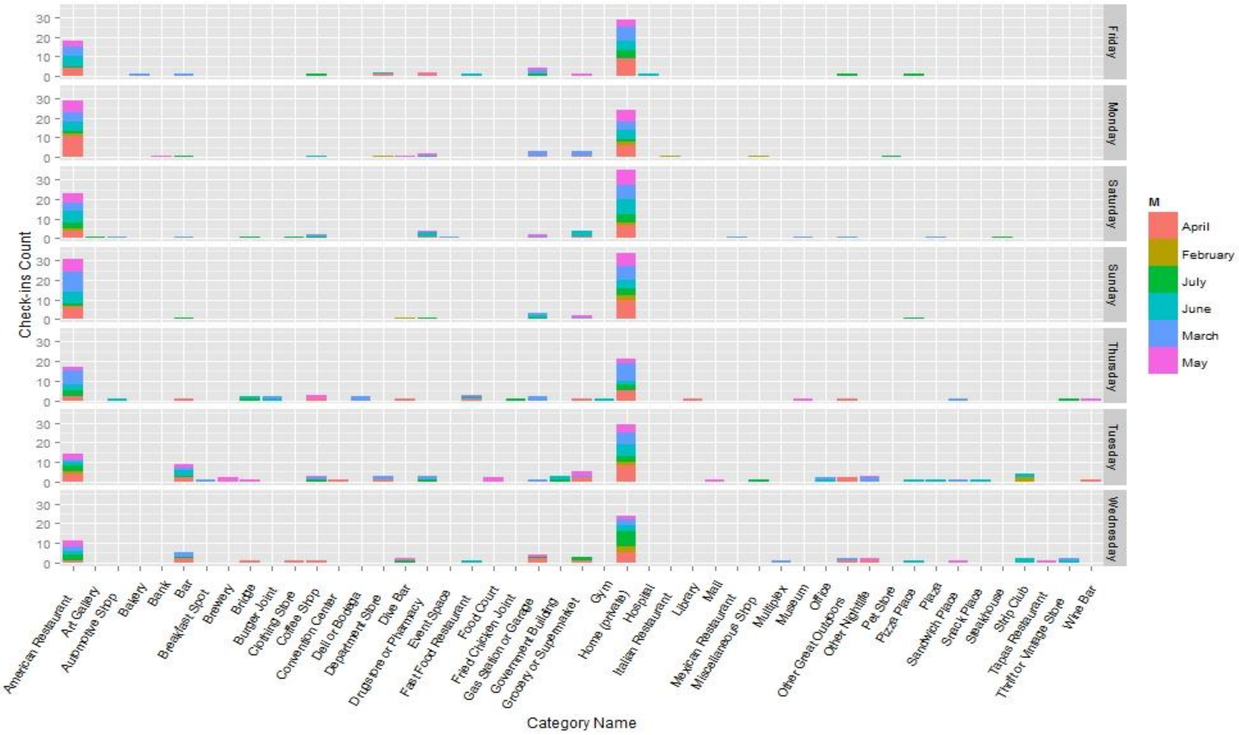


Figure 5. The frequent user's check-ins count in different categories of venues, classified by day and grouped by month.

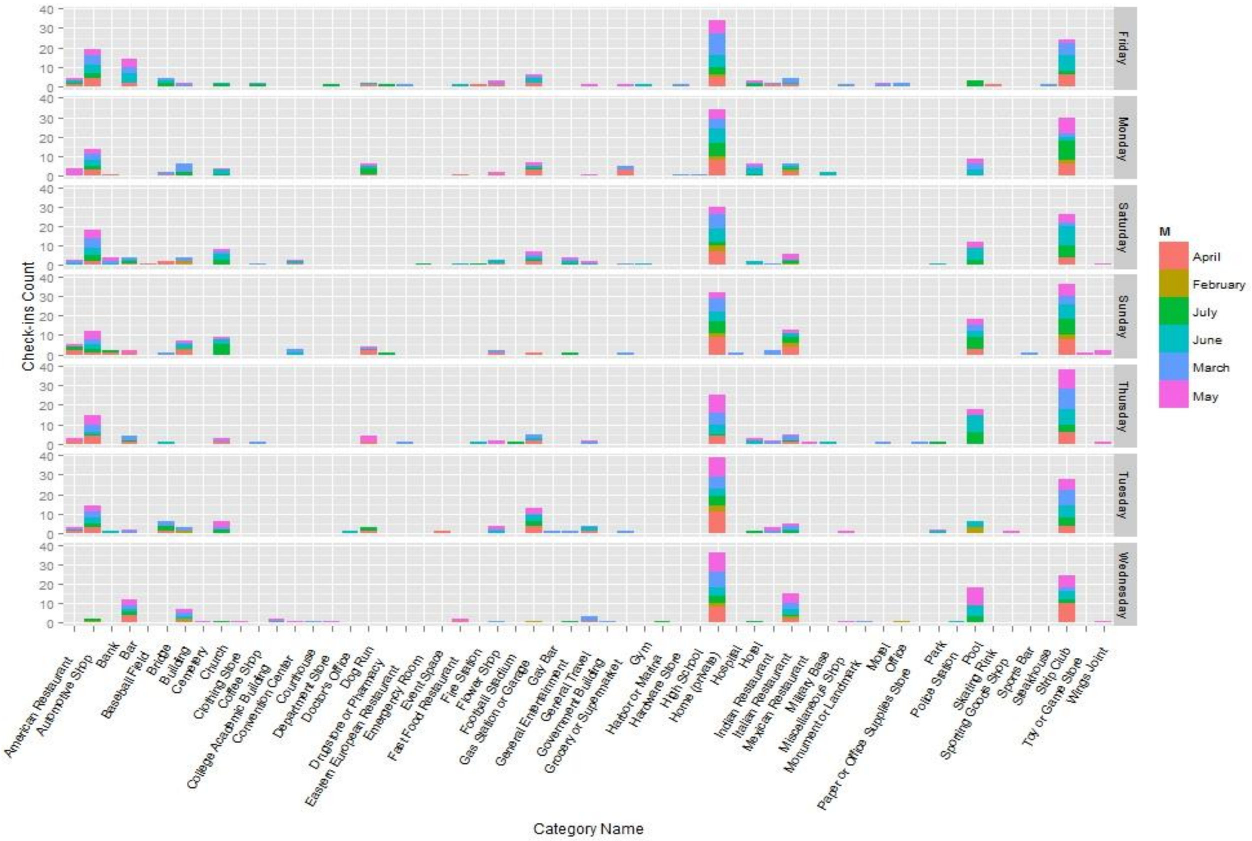


Figure 6. The hyper-active user's check-ins count in different categories of venues, classified by day and grouped by month.

An Evaluation of Semantically Enriched Spatial Data Infrastructure

Hamon Barros Henriques
Computer and Systems Depart.
Federal Univ. of Campina Grande
Campina Grande, Brazil
hamon@copin.ufcg.edu.br

Fabio Gomes de Andrade
Computer and Systems Depart.
Federal Institute of Paraíba
Cajazeiras, Brazil
fabio@ifpb.edu.br

Cláudio de Souza Baptista
Computer and Systems Depart.
Federal Univ. of Campina Grande
Campina Grande, Brazil
baptista@dsc.ufcg.edu.br

***Abstract*— Spatial Data Infrastructures (SDI) have become an important solution for easing the interoperability of geographic data offered by different organizations. An important challenge that must be overcome by such infrastructures consists in allowing their users to easily locating the available data. Presently, this task is implemented by means of catalog services, which still have important limitations that prevent effective data retrieval. Due to those limitations, many research works have been developed to improve information retrieval in SDIs. One of such works is Semantically Enabled Spatial Data Infrastructures (SESDI), which is a framework that uses a model-based on information at feature type level and ontologies. The first results obtained during the experimental evaluation of SESDI showed that it improved the quality of several kinds of queries concerning geographic data. Nevertheless, a deeper evaluation, besides the comparison to catalog services provided by other infrastructures, was still necessary. Aiming at meeting this need, this paper describes an experiment carried out in order to deepen that evaluation. In this experiment, the performance of SESDI was compared with catalog services offered by other two infrastructures. The results obtained from the new experiments showed the viability of the solution used to implement the framework.**

Keywords-spatial data infrastructure; catalog service; spatial database; GIS.

I. Introduction

Recently, Spatial Data Infrastructures (SDI) have become popular as an important solution for allowing the interoperability of geographic data supplied by different organizations. In order to achieve this interoperability, these infrastructures are created based on a set of norms and standards that must be adopted by all of their components.

In this standardization scenario, the standards specified by the Open Geospatial Consortium (OGC) [1] have played a key role. Important examples of these standards include the services that give support for accessing geographic data, such as Web Map Service (WMS) [2] and Web Feature Service (WFS) [3]. These services allow the clients to have access to geographic data offered by a given provider using a standard interface, with no need to be aware of the details about its storage. Each instance of these services gives access to a set of feature types. On the other hand, each feature type represents a layer present in the dataset offered by a provider. Since their proposal, the OGC web services have played a key role for the development of spatial data structures, being used in the implementation of many of the current infrastructures.

Besides interoperability, an important challenge which must be overcome by SDIs consists in allowing the clients to easily locating the available data and services. Presently, this problem is usually solved by the implementation of catalog services. In a catalog service, geographic data providers must supply a metadata containing detailed information about their dataset. On the other hand, clients of a SDI may use this service in order to find out the geographic data they are interested in.

Although having led to important advances, the present catalog services still have some limitations. One of these limitations is that they solve their queries solely based on the information contained in the metadata records created by the geographic data providers. Since these records normally describe the datasets offered by a service as a whole, they usually bring little or even no information concerning the spatial, thematic and temporal characteristics of each feature type, which constrains the information retrieval process. Another important limitation is that most of the present catalogs solve queries with thematic constraints based on keywords only. This characteristic causes the catalog to discard, during the information retrieval process, relevant resources that are described with terms related to the theme defined in the query, which reduces the recall of these queries.

Throughout the years, many research works have been proposed with the objective of overcoming these limitations. One such work is Semantically Enabled Spatial Data Infrastructures (SESDI) [4]. SESDI is a framework that uses a model-based on metadata at feature type level and ontologies to improve the retrieval of geographic data offered by an SDI. During the evaluation of this framework, the results achieved by its search engine were compared with the results achieved by similar queries submitted to the catalog service offered by a present SDI. This validation, which used the North-American SDI (NSDI) [5] as case study, showed that SESDI improved the quality of several kinds of queries. Nevertheless, a deeper evaluation, besides

ISBN: 978-1-61208-326-1

the comparison to other available catalog services, was still necessary.

Aiming at meeting this need, this paper describes an experiment carried out to deepen the evaluation of SESDI. In this experiment, the performance of SESDI was compared with the catalog services offered by other two infrastructures, namely the Canadian Spatial Data Infrastructure [6] and the Global Earth Observation System of Systems (GEOSS) [7]. During the experiment, each solution was evaluated based on four query types: spatial, thematic, temporal and global. Besides allowing a deeper evaluation of SESDI, this paper offers as contribution an analysis of the performance of the studied catalog services, evaluating their quality at solving several kinds of queries involving spatial data.

The remaining of this paper is organized as follows. In Section II, we approach related works. Section III gives an overview of the two approaches studied in this work. Section IV focuses on the experiment design, describing the dependent variables, the response variables, the experimental units, the research hypotheses and the data collection process. Section V describes the results achieved with the experiment. Finally, Section VI concludes the paper, and discusses further work to be undertaken.

II. Related Works

Recently, many works have been proposed with the objective of improving the retrieval of geographic data and services from spatial data infrastructures and geographic portals. Some of these works are concerned with the retrieval of services. Stock et al. [8] developed a solution in which the services are retrieved from information defined in a feature type catalog that defines the structure and the relationships among the feature types offered by the SDI. Lutz et al. [9] associated the output parameters of each service to concepts defined in application ontologies to improve information retrieval in disaster management applications. In another work, Lutz [10] used first order logic to generate a signature which describes the semantic of the services offered by an SDI. In the approach developed by Lemmens et al. [11], services were classified and retrieved according to a geospatial web services ontology. Other proposed solutions [12][13] described the available services as tasks. Li et al. [14] proposed a solution which expands the terms defined in the query of a client in order to improve the retrieval of geographic web services that offer data about the Arctic region. The disadvantage of all these works is that they do not take into account the information offered at feature type level during the resolution of queries.

Some of the developed works manage to solve queries based on information at feature type level. Janowicz et al. [15] developed a solution which used a semantic similarity measure based on ontologies to retrieve geographic data. In another work, Zhang et al. [16] proposed a solution that resolves queries for feature types using specific ontologies related to four types of dimension: location, theme, geometry and properties. Li et al. [17], in turn, implemented a solution that resolves thematic queries at feature type level using a keyword-based approach, which constrains the quality of this kind of query. A disadvantage of all these studies is that they do not approach the resolution of temporal queries, which are important for many queries involving geographic data.

Finally, some proposed works deal with the retrieval of geographic metadata, and do not focus on a specific kind of resource. Smits and Friis-Christensen [18] developed a solution which uses a multilingual thesaurus to describe the information offered in the catalog of an infrastructure. Athanasis et al. [19], in turn, developed a solution that describes the resources offered by an SDI by means of metadata based on a series of ontologies. Chen et al. [20] used OWL-S to describe the semantics of data offered through geographic web services. Despite their importance and relevance, these works cannot solve queries based on information at feature type level. Also, these solutions do not support the resolution of temporal queries.

III. CatAlog Sevices versus SESDI

This section provides an overview of the two approaches that were compared during the experiment. First, we show how the queries are solved by the current catalog services. After that, we show the approach used in the implementation of SESDI.

A. *Catalog Service*

Currently, geographic data providers use the catalog service to advertise their resources. For this, they create a new metadata record into the catalog service. During this process, they supply metadata that describe their datasets, according to the geographic metadata standard adopted by the SDI. Some examples of these metadata include the title of the dataset, the coordinate systems and the projection used in the production of data, and the URL from which the service can be invoked. The metadata provided by each metadata record are used by the catalog service during the query resolution process. Since geographic data providers normally use a single metadata record to describe their datasets as a whole, current catalog services have limitations to solve spatial, thematic, temporal and global queries.

Regarding spatial queries, the main limitation is that data providers normally define a single bounding-box to represent their datasets. Then, during the searching process, the catalog selects all the records whose bounding-box intersects (or contains) the geographic region defined in the query. Nevertheless, it is possible to notice that sometimes the feature types provided in a dataset cover different regions, which causes some limitations to the searching process. To understand these limitations, suppose the case depicted in Figure 1, which describes a dataset about flooding. In that figure, the feature types present in the dataset are represented as green rectangles, while the blue rectangle represents the spatial, thematic and temporal information provided in the metadata record that describes the dataset. Also, suppose B1 and B2 are two bounding-boxes covering two different regions that do not overlap each other. In Figure 1, the geographic extent provided in the record is B1, which represents the region that is covered by most of the feature types of the dataset. Then, if a user poses a query looking for maps about the region B2, the catalog service discards the record depicted in that figure, even its dataset has a feature type that satisfies the search criteria. Moreover, if a user poses a query for maps about the region B1, the catalog returns the record. After that, the user needs to access the service to identify, among the entire feature types of the dataset, just the ones that are relevant to his/her query. This process can be quite tedious and tiring for the user, since a query may return a large volume of records and each service, in turn, may offer a large number of feature types. The limitations regarding spatial queries can be extended to temporal queries. Moreover, in temporal queries these limitations are even bigger, since many providers do not provided the metadata about the temporal extent of the dataset.

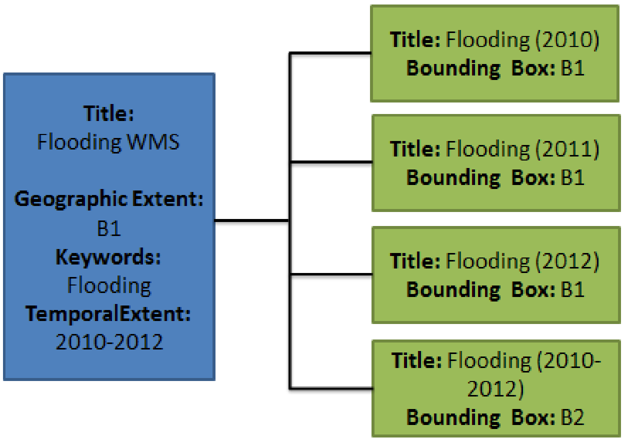

Title: Flooding WMS
Geographic Extent: B1
Keywords: Flooding
TemporalExtent: 2010-2012
Title: Flooding (2010) Bounding Box: B1
Title: Flooding (2011) Bounding Box: B1
Title: Flooding (2012) Bounding Box: B1
Title: Flooding (2010-2012) Bounding Box: B2

Figure 1. Example of a metadata record.

Regarding thematic queries, the main limitation is that catalog normally solves queries based on keywords. During the searching process, the catalog selects the records that contain in their description the keyword used in the query. Then, relevant records that contain in their description keywords that are related to the theme requested in the query can be discarded during this process. For example, suppose the case depicted in Figure 1 again. If a user poses a query for maps about *disasters*, the record showed in the figure is not selected by the catalog service, even its dataset offers several feature types that are relevant to the user. Finally, global queries are even more difficult to solve by using the current catalogs, since they present the limitations concerning all the dimensions used in the query.

B. SESDI

SESDI is a framework proposed with the objective of easing the retrieval of geographic data offered by SDIs, and aims to overcome some of the limitations of the present catalog services. In order to achieve this goal, it solves the queries of the clients based on a model that adapts classic Information Retrieval (IR) techniques to the domain of geographic data. In this model, the services that offer access to geographic data, such as WMS and WFS, are described as a set of feature types, in the same way as the documents are described as a set of keywords in the classic information retrieval models. Another important characteristic of SESDI is that it uses ontologies to describe the semantic of the feature types offered by the service, in order to improve the quality of the queries with thematic constraints. Finally, the framework proposes a search engine that explores the spatial, thematic and temporal relationships among the feature types offered by a service in order to generate the results retrieved in their queries.

In order to implement its model, the SESDI identifies the spatial, thematic and temporal characteristics of each feature type present in the dataset offered by the services registered in the catalog service of the infrastructure. To perform this task, it processes the information of each metadata record in order to retrieve information concerning the services offered by the SDI. After that, it invokes the *GetCapabilities* operation of each identified service in order to retrieve its capabilities document. The objective of this stage is to retrieve more detailed information about its feature types.

At the end of these stages, the framework processes the information contained in the metadata record and in the capabilities document to identify the spatial, thematic and temporal characteristics of each feature type. This entire process is called tagging and is split into three stages: spatial tagging, thematic tagging and temporal tagging. It is important to take in mind that the tagging process is executed for each feature type identified by the framework. The following paragraphs describe the tagging process. However, more detailed information about this process can be found in [4].

In the spatial tagging, the SESDI identifies the geographic region covered by the feature type. That information, which is represented as a bounding box, is obtained from the feature type description in the capabilities document retrieved from the service.

In the thematic tagging, the framework tries to identify the semantics of the data provided by the feature type. It accomplishes that task by relating the feature types to concepts defined in ontologies. In the current version, the SESDI uses a set of ontologies about several application domains. During the thematic tagging, the framework processes the keywords list provided for the feature type in the capabilities document of the service. If no keywords are provided for the feature type, its title is used as the input for that process. Then, the SESDI matches each keyword (or the title of the feature type) to the names of the concepts used in its ontologies. Whenever a match is identified, the framework generates a tag associating the matched concept to the feature type. When a matching cannot be identified, the SESDI poses a query in the Wikipedia and tries to retrieve a page related to the keyword. If more than one page is retrieved, a ranking for each page is generated by using techniques of classic information retrieval, and the page with the highest ranking value is selected. After that, the framework matches the title of the selected page, as well the name of each category used in the page description, to the concepts used in its ontologies. The thematic tags generated during this process are associated to the feature type description and stored in a database. If no matches are identified during the tagging process, the framework does not generate any tag to the feature type.

Finally, the temporal tagging consists of identifying the time period of the feature type. Since the capabilities document does not provided metadata specific to describe temporal information, the SESDI processes the title and the text description of the feature type in order to find temporal expressions that provides that information. When one or more temporal expressions are found, their values are converted into a time interval. On the other hand, if the framework cannot identify any temporal expression, it assumes that the time interval of the feature type is the same defined in the metadata record. Furthermore, if no value is provided for temporal extent in the metadata record, the value null is used as the time interval of the feature type. After the tagging process, the metadata generated by SESDI are stored in a local database, along with some description about the feature type and its respective service. Figure 2 shows the metadata generated by SESDI for the feature types depicted in Figure 1 after the tagging process.

A key difference between SESDI and the present catalog services is that SESDI solves queries based on the metadata identified for each feature type during the tagging process. Another important difference is that SESDI returns to the users only the feature types that satisfy the search criteria, so clients do not need to access the service to identify the feature types of his/her interest. For example, if a user poses a spatial query for maps about the region B2, the framework is able to identify that the feature type "*Flooding (2010-2012)*" is relevant for the query. Moreover, if a user poses a query for maps about the region B1, the framework returns only the feature types that satisfy the constraint defined in the query. Another important difference is that the SESDI uses thematic tags associated to concepts defined the ontologies. Then, when a user poses a query for a specific theme, the framework is able to return as the feature types that are tagged exactly with the theme used in the query as the ones that are tagged with a concept related to it.

So, the study described in this paper was intended to carry out a deeper evaluation of the performance of these two kinds of approach. This approach was based on an empirical experiment, where we defined the controlled variables, which are the variables subject to adjustments before the execution of the experiment, and the dependent variables, which result from the experiment. Besides, with these variables, we elaborated the hypotheses to be verified. The next sections describe, respectively, the experiment design and the evaluation of the results.

IV. EXPERIMENT DESIGN

The experiment followed the same strategy of the evaluation used in [4], which employed recall and precision metrics to evaluate the performance of SESDI with respect to catalog services of the Canadian SDI and of the GEOSS. For each kind of query (spatial, temporal, semantic and global), the performance was measured according to the recall (number of relevant results) and precision (quality of the relevance) of the results. On the other hand, at the service level, recall and precision were evaluated with basis on the number of retrieved services.

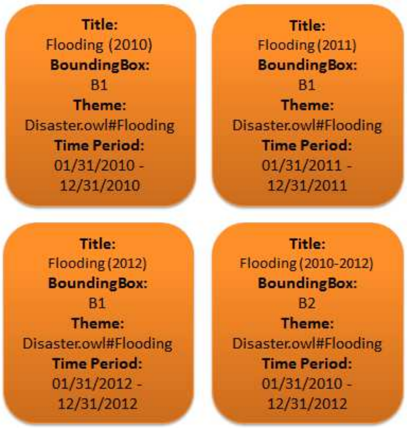


Figure 2. Example of feature type metadata generate by SESDI.

The kinds of queries used in the research are classified as purely spatial, purely semantic, purely temporal and global. The purely spatial queries are intended to search for maps that intersect, partially or entirely, a certain location of interest to the user (e.g., which cities are crossed by the São Francisco river?). The purely semantic queries, in turn, are intended to retrieve maps that describe geographic data about a certain theme (subject) (e.g., which maps refer to beaches?). The purely temporal queries have the objective of retrieving the maps that refer to a certain time interval (e.g., *which maps refer to the decade of 1950*?). Finally, the global queries retrieve maps that deal with a certain theme in a certain location and in a time interval, that is, these queries meet more than one constraint type (e.g., *which maps refer to beaches in Brazil in the decade of 1950*?).

The controlled variables used in the experiment were: kind of query used (purely spatial, purely temporal, purely semantic or global) and the used tools (SESDI and the catalog services of the Canadian SDI and of the GEOSS). On the other hand, the dependent variables used to compare the performance of SESDI with respect to the studied SDIs were recall and precision. These are the main metrics for evaluation of the performance of the information retrieval system. The recall is measured as the ratio between the number of relevant resources retrieved and the number of relevant resources that exist in the system. The precision is measured as the ratio between the number of relevant resources retrieved and the total number of resources retrieved.

The experimental units were composed by the comparisons of SESDI with the Canadian SDI and the GEOSS. For each of them, twenty queries of each type were performed [21]. For each round of testing, the query that was held in the SESDI was held too in the SDIs studied. Moreover, during the comparison between the SESDI and the SDIs, the queries were held on the same set of metadata.

Before the formulation of the hypotheses, the concepts of independent variables and response variables were formalized, in order to provide a better visualization of the presented information. The independent variable "query type" was formalized by the symbols *e*, *s*, *t* and *g*, for the spatial, semantic, temporal and global levels, respectively, while the independent variable "used tool" was formalized by the symbols *sd*, *c* and *g*, representing SESDI, the Canadian SDI and the GEOSS SDI, respectively. In order to represent the response variables "precision" and "recall", the following functions were formalized:

- *PFT*: represents the precision at feature type level, which evaluates the quality of the feature types retrieved from each approach;
- *PS*: represents the precision at service level, which evaluates the quality of the services retrieved from each approach;
- *CF*: represents the recall at feature type level, which evaluates the quantity of relevant feature types retrieved from each approach; and
- *CS*: represents the recall at service level, which evaluates quantity of relevant services retrieved from each approach.

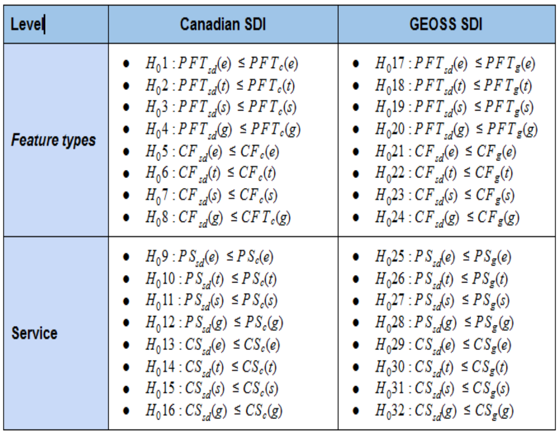
The objective of this study was to evaluate the performance of SESDI with respect to the catalog services provided by some existing SDIs. The null hypotheses formalized to describe the comparisons between SESDI and the SDIs used as case studies are depicted in Table I. The first column describes if the hypothesis represent a query at level of service or feature type. The second and third columns represents, respectively, the null hypotheses used during the comparison between SESDI and the catalog service provided by the Canadian SDI and GEOSS. In Table I, each hypothesis represents a comparison between the performance of SESDI and the catalog service provided by a SDI. These hypotheses were formulated based on the metrics presented above (*PFT*, *PS*, *CF* and *CS*). For example, the hypothesis H_01assumes that the precision at feature type level in the spatial queries held in SESDI is less or equal than the precision at feature types level in the spatial queries held in Canadian SDI. For each hypothesis defined in Table I, we performed statistical tests to try to refute it. The results of statistical tests are showed in Tables II and III.

The process of collecting the research data occurred in a semi-automatic and individual manner for each one of the queries.

The first step was the generation of the baseline. In this stage, the relevant results to be retrieved in one of the query types were obtained. For this, the records which should be returned by a query were previously selected, generating a baseline that was used to compare the results obtained from each approach. Once generated, the baseline was stored in a file. Next, we present the criteria used for generation of the baseline for each query type:

- **Purely spatial query:** the baseline for the feature type level was composed of all the layers whose bounding-box intersected the geographic region defined in the query. On the other hand, for the service level, the baseline was composed of all the services that offered at least one layer present in the baseline for the feature type level;

TABLE I - TABLE OF FORMALIZED NULL HYPOTHESES



Level	Canadian SDI	GEOSS SDI
Feature types	• $H_01 : PFT_{sd}(e) \leq PFT_c(e)$ • $H_02 : PFT_{sd}(t) \leq PFT_c(t)$ • $H_03 : PFT_{sd}(s) \leq PFT_c(s)$ • $H_04 : PFT_{sd}(g) \leq PFT_c(g)$ • $H_05 : CF_{sd}(e) \leq CF_c(e)$ • $H_06 : CF_{sd}(t) \leq CF_c(t)$ • $H_07 : CF_{sd}(s) \leq CF_c(s)$ • $H_08 : CF_{sd}(g) \leq CFT_c(g)$	• $H_017 : PFT_{sd}(e) \leq PFT_g(e)$ • $H_018 : PFT_{sd}(t) \leq PFT_g(t)$ • $H_019 : PFT_{sd}(s) \leq PFT_g(s)$ • $H_020 : PFT_{sd}(g) \leq PFT_g(g)$ • $H_021 : CF_{sd}(e) \leq CF_g(e)$ • $H_022 : CF_{sd}(t) \leq CF_g(t)$ • $H_023 : CF_{sd}(s) \leq CF_g(s)$ • $H_024 : CF_{sd}(g) \leq CF_g(g)$
Service	• $H_09 : PS_{sd}(e) \leq PS_c(e)$ • $H_010 : PS_{sd}(t) \leq PS_c(t)$ • $H_011 : PS_{sd}(s) \leq PS_c(s)$ • $H_012 : PS_{sd}(g) \leq PS_c(g)$ • $H_013 : CS_{sd}(e) \leq CS_c(e)$ • $H_014 : CS_{sd}(t) \leq CS_c(t)$ • $H_015 : CS_{sd}(s) \leq CS_c(s)$ • $H_016 : CS_{sd}(g) \leq CS_c(g)$	• $H_025 : PS_{sd}(e) \leq PS_g(e)$ • $H_026 : PS_{sd}(t) \leq PS_g(t)$ • $H_027 : PS_{sd}(s) \leq PS_g(s)$ • $H_028 : PS_{sd}(g) \leq PS_g(g)$ • $H_029 : CS_{sd}(e) \leq CS_g(e)$ • $H_030 : CS_{sd}(t) \leq CS_g(t)$ • $H_031 : CS_{sd}(s) \leq CS_g(s)$ • $H_032 : CS_{sd}(g) \leq CS_g(g)$

- **Purely temporal query:** for the feature type level, the baseline was composed of all the layers whose temporal extension intersected the time interval defined in the query. On the other hand, for the service level, the baseline was composed of all the services that offered at least one layer present in the baseline for the feature type level;
- **Purely semantic query:** for the feature type level, the baseline was composed of all the layers that offered data about the theme used in the request, or a theme related to it. On the other hand, for the service level, the baseline was composed of all the services that offered at least one layer present in the baseline for the feature type level;
- **Global query:** for the feature type level, we considered all the layers that met the three constraints (spatial, temporal, and semantic) defined in the request. On the other hand, for the service level, the baseline was composed of all the services that offered at least one layer present in the baseline for the feature type level.

The second step was the execution of the queries. In this stage, the query was formulated according to its criteria and performed both in SESDI and in the catalog service of the SDI.

Finally, in the last step, we compared the results obtained from the two approaches. During this stage, we generated an output file containing information about the experiment. This file stored information about the number of services and feature types in each baseline. Moreover, for each approach, the file stored the number of services retrieved, the number of feature types retrieved, the number of relevant services retrieved and the number of relevant feature types retrieved. Then, we used the information stored in that file for the computation of the precision and recall metrics.

In order to make the analysis of results easier, a tool was developed to compute the data concerning the dependent variables (precision and recall) of this research, consolidating them into a single file. This way, the consolidated data can be observed in the files related to the comparison between SESDI and the Canadian SDI [22] and between SESDI and the GEOSS SDI [23].

V. Results of the Experimental Evaluation

After collecting the data, it was necessary to employ the hypotheses tests to infer about the hypotheses cited in previous section. To accomplish this task, we used a methodology proposed by Wohlin [24]. According to this method, we firstly verified the normality, aiming to check whether the data come from a normal distribution. Next, we collected the homoscedasticity of the data in order to tell whether there is variance or not in the data. Both normality and homoscedasticity allow one to decide about the use of parametric or non-parametric tests. Since every statistical test needs a null hypothesis to infer about it, the tests used here result in a p-value (probability value) that, depending on its value, may deny or not the hypothesis of the test according to the significance level adopted.

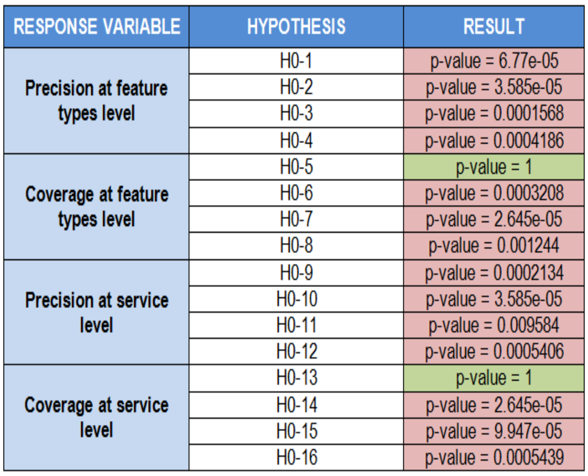
In order to verify the normality, we used the Shapiro-Wilk, Anderson-Darling, Skewness and Kurtosis tests [24], with significance level of five percent, besides the QQ-plot graphics. Moreover, in the comparison of SESDI with the Canadian SDI and the GEOSS, the data are not from a normal distribution. Since most tests point that the data involving controlled variables with the dependent variables are not normal, a few data were pointed as normal in some tests, but the QQ-plot graphic concerning them allowed us to see their non-normality. The results of the tests and the QQ-plot graphics can be seen in [25].

Since the collected data are not normal, the Levene test [24] was used to perform the verification of the homoscedasticity aspects with a significance level of five percent. The results of the tests related to the data of the comparisons between the SESDI and the catalog services of the Canadian SDI and of the GEOSS can be seen in [25].

As we mentioned before, the data were not retrieved from a normal population. So, we used a non-parametric test to infer about the experiment hypotheses. Since the data were collected in a dependent manner, we needed to use a non-parametric test of the paired samples. Besides, the hypotheses of the experiment compared to data groups. For this reason, we chose the Mann–Whitney with significance level of five percent.

The results of the test for each hypothesis listed in Table I, related to the performance comparison between SESDI and the Canadian SDI, are presented in Table II. Such results are coded in two colors: green, when the null hypothesis was not refuted (p-value is above the significance level adopted - 5%) and red, when the null hypothesis was refuted (p-value is smaller than the significance level adopted - 5%). Table II shows that most of the hypotheses related to comparison between SESDI and the Canadian SDI were refuted. These results lead us to conclude that SESDI had better performance for most cases, since the null hypotheses presented assume the inferiority or equality of the response variables for all query types. The table shows that the hypotheses *H0-5* and *H0-13* were not refuted, which means that we cannot state that SESDI improves the recall for spatial queries at service and feature type levels.

Table II - Results of the Mann-Whitney Tests Applied to the Comparison of SESDI with the Canadian SDI

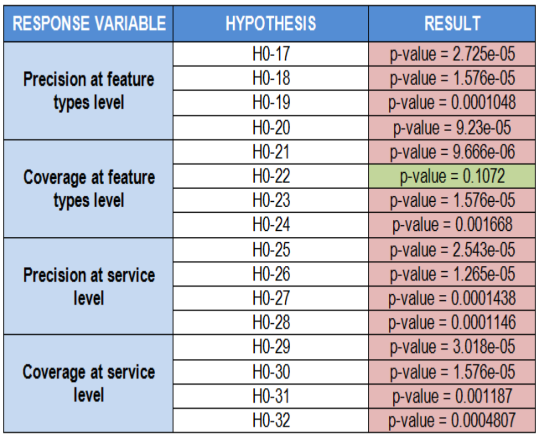


RESPONSE VARIABLE	HYPOTHESIS	RESULT
Precision at feature types level	H0-1	p-value = 6.77e-05
	H0-2	p-value = 3.585e-05
	H0-3	p-value = 0.0001568
	H0-4	p-value = 0.0004186
Coverage at feature types level	H0-5	p-value = 1
	H0-6	p-value = 0.0003208
	H0-7	p-value = 2.645e-05
	H0-8	p-value = 0.001244
Precision at service level	H0-9	p-value = 0.0002134
	H0-10	p-value = 3.585e-05
	H0-11	p-value = 0.009584
	H0-12	p-value = 0.0005406
Coverage at service level	H0-13	p-value = 1
	H0-14	p-value = 2.645e-05
	H0-15	p-value = 9.947e-05
	H0-16	p-value = 0.0005439

The results of the test for each hypothesis listed in Table I, related to the performance comparison between the SESDI and the GEOSS SDI, are presented in Table III. In that table, it is possible to notice also that most of the hypotheses related to the performance comparison between the SESDI and the GEOSS SDI were refuted, which allows us to conclude that the performance of the SESDI was superior to that of the GEOSS SDI for most cases. The only hypothesis that was not refuted (*H0-22*) shows that there is no statistical significance to state that SESDI improved recall of temporal queries at feature type level.

To better illustrate the results presented above, the returns a semantic query used in the experiment are shown below. The consultation aimed to find maps for "boundaries" and obtained a precision level of feature type 100% in the SESDI, i.e., all relevant services were recovered, while the IDE obtained at a precision level of 11% of feature type.

Table III - Results of the Mann-Whitney Tests Applied to the Comparison Between SESDI and GEOSS



RESPONSE VARIABLE	HYPOTHESIS	RESULT
Precision at feature types level	H0-17	p-value = 2.725e-05
	H0-18	p-value = 1.576e-05
	H0-19	p-value = 0.0001048
	H0-20	p-value = 9.23e-05
Coverage at feature types level	H0-21	p-value = 9.666e-06
	H0-22	p-value = 0.1072
	H0-23	p-value = 1.576e-05
	H0-24	p-value = 0.001668
Precision at service level	H0-25	p-value = 2.543e-05
	H0-26	p-value = 1.265e-05
	H0-27	p-value = 0.0001438
	H0-28	p-value = 0.0001146
Coverage at service level	H0-29	p-value = 3.018e-05
	H0-30	p-value = 1.576e-05
	H0-31	p-value = 0.001187
	H0-32	p-value = 0.0004807

VI. CONCLUSION AND FUTURE WORKS

This research was intended to perform a deeper evaluation of the performance of the SESDI framework during the retrieval of geographic data offered by SDIs. In order to perform this evaluation, the performance of SESDI was compared with the performance of two catalog services presently offered by two infrastructures: The Canadian SDI and the GEOSS SDI. Faced with the results presented above, one can conclude, with a significance level of five percent, that:

1) For all types of queries, SESDI had better precision than the Canadian SDI for most of the queries executed at the feature type level;
2) For the global, purely temporal and purely semantic query types, SESDI had a better recall at feature type level, compared to the Canadian SDI. As for the purely spatial queries, there was not statistical significance to state which approach had better performance;
3) For all types of query, SESDI had better precision in the service level than the Canadian SDI;
4) For the global, purely temporal and purely semantic query types, SESDI had a better recall at service level than the Canadian SDI. As for the purely spatial queries, there was not statistical significance to assert which approach had better performance;
5) For all types of query, SESDI had better precision at the feature type level than the GEOSS SDI;
6) For the purely spatial, purely semantic and global query types, SESDI had a better recall at the feature type level than the GEOSS SDI. As for the purely spatial queries, there was not statistical significance to state which approach had better performance;
7) For all types of query, SESDI had better precision at service level than the GEOSS SDI;
8) For all types of query, SESDI had better recall at service level than the GEOSS SDI.

Based on the results, we could conclude that the solution used to implement SESDI is viable, since it improved recall and precision for most queries used during experimental evaluation. As a suggestion for future work, SESDI could be compared with other catalog services. The results achieved with such experiments could add still more value to the object of study of this research.

Another suggestion for future works would be the implementation of the universalization of the language used by SESDI. With this, it would be possible to compare SDIs with languages other than English.

REFERENCES

[1] Lupp, M. (2008). Open geospatial consortium. In Shekhar, S. and Xiong, H., editors, Encyclopedia of GIS, page 815. Springer.

[2] Open Geospatial Consortium, “Ogc web map service interface,” OGC 03-109r1, Version 1.3.0, January 2004.

[3] ——, “Web feature service implementation specification,” OGC 04-094, Version 1.1.0, May 2005.

[4] F. G. de Andrade, C. de S. Baptista, and C. A. Davis Jr, “Improving geographic information retrieval in spatial data infrastructures.” GeoInformatica, Jan. 2014, doi: 10.1007/s10707-014-0202-x.

[5] http:// http://www.fgdc.gov/nsdi/nsdi.html [accessed: 2014-01-20]

[6] http://geoconnections.nrcan.gc.ca/[accessed: 2014-01-20]

[7] http://www.earthobservations.org/geoss.shtml[accessed: 2014-01-20]

[8] K. M. Stock et al., “A semantic registry using a feature type catalogue instead of ontologies to support spatial data infrastructures.” International Journal of Geographical Information Science, vol. 24, no. 2, Feb. 2010, pp. 231–252.

[9] M. Lutz, J. Sprado, E. Klien, C. Schubert, and I. Christ, “Overcomingsemantic heterogeneity in spatial data infrastructures.” Computers Geosciences, vol. 35, no. 4, Apr. 2009, pp. 739–752.

[10] M. Lutz, “Ontology-based descriptions for semantic discovery and composition of geoprocessing services.” GeoInformatica, vol. 11, no. 1, Mar. 2007, pp. 1–36.

[11] R. Lemmens, R. A. de By, M. Gould, A. Wytzisk, C. Granell, and P. van Oosterom, “Enhancing geo-service chaining through deep service descriptions,” T. GIS, vol. 11, no. 6, Dec. 2007, pp. 849–871.

[12] M. Molina and S. Bayarri, “A multinational sdi-based system to facilitate disaster risk management in the andean community.” Computers Geosciences, vol. 37, no. 9, Sep. 2011, pp. 1501–1510.

[13] N. Wiegand and C. García, “A task-based ontology approach to automate geospatial data retrieval,” T. GIS, vol. 11, no. 3, Jun. 2007, pp. 355–376.

[14] W. Li, et al., “Semantic-based web service discovery and chaining for building an arctic spatial data infrastructure.” Computers Geosciences, vol. 37, no. 11, Nov. 2011, pp. 1752–1762.

[15] K. Janowicz, M. Wilkes, and M. Lutz, “Similarity-based information retrieval and its role within spatial data infrastructures.” in GIScience, ser. Lecture Notes in Computer Science, T. J. Cova, H. J. Miller, K. Beard, A. U. Frank, and M. F. Goodchild, Eds., vol. 5266. Springer, Sep. 2008, pp. 151–167.

[16] C. Zhang, T. Zhao, W. Li, and J. P. Osleeb, “Towards logic-based geospatial feature discovery and integration using web feature service and geospatial semantic web.” International Journal of Geographical Information Science, vol. 24, no. 6, Jun. 2010, pp. 903–923.

[17] Z. Li, C. P. Yang, H. Wu, W. Li, and L. Miao, “An optimized framework for seamlessly integrating ogc web services to support geospatial sciences.” International Journal of Geographical Information Science, vol. 25, no. 4, Apr. 2011, pp. 595–613.

[18] P. C. Smits and A. Friis-Christensen, “Resource discovery in a european spatial data infrastructure.” IEEE Trans. Knowl. Data Eng., vol. 19, no. 1, Jan. 2007, pp. 85–95.

[19] N. Athanasis, K. Kalabokidis, M. Vaitis, and N. Soulakellis, “Towards a semantics-based approach in the development of geographic portals,” Comp. Geosci, vol. 35, Feb. 2009, pp. 301–308.

[20] N. Chen, Z. Chen, C. Hu, and L. Di, “A capability matching and ontology reasoning method for high precision ogc web service discovery.” Int. J. Digital Earth, vol. 4, no. 6, Jun. 2011, pp. 449–470.

[21] http://zip.net/bcmbc2[accessed: 2014-01-20]

[22] http://zip.net/bkmbzd[accessed: 2014-01-20]

[23] http://zip.net/bkmbzc[accessed: 2014-01-20]

[24] C. Wohlin et al., Experimentation in Software Engineering. Heidelberg, Berlin: Springer-Verlag, 2012.

[25] http://zip.net/bll9Wz[accessed: 2014-01-20]

Integrating GIS Visualization Tools for Ecosystem Management

Keith L. Lee, David Stotts
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
{lee,stotts}@cs.unc.edu

Jennifer A. Moore Myers, Emrys Treasure, Robert L. Herring, Steve G. McNulty
USDA Forest Service, Southern Research Station
Eastern Forest Environmental Threat Assessment Center
Raleigh, NC, USA
{jamyers, eatreasu,rlherrin,steve_mcnulty}@ncsu.edu

***Abstract*—This paper presents a targeted view into the current development stage of the 3.X version of TACCIMO, a web-based application delivering climate change science to the forest management and planning community. The interactive TACCIMO GIS Viewer delivers custom visualizations and reports combining climate change projections with peer-reviewed literature and planning language. We present our work integrating externally hosted, heterogeneous data sources into the TACCIMO GIS Viewer, including our efforts to accommodate minimal changes to existing source code, along with advanced geospatial analysis methods using integrated climate change data. Our real world case studies demonstrate the contexts motivating our work for users assessing, managing, and monitoring forest resources within the conterminous United States.**

Keywords-ecosystem management; forestry; Geographic Information Systems; geospatial map service; web service; climate change.

I. INTRODUCTION

Climate change science plays an important role in forest planning and management [1]. Land managers pose seemingly straight-forward questions, e.g., “what tree species do we need to plant,” whose answers depend on a myriad of parameters including climate change science. Scientists and researchers develop tools used by land managers and policymakers to improve their understanding of climate change and its impact on forests, rangelands, and urban areas [2]. Land managers and policymakers then use this information in developing policies and management techniques to sustain and improve ecosystems of interest [3][4].

Given the ever-increasing volume of scientific knowledge and tools regarding climate change and forest ecosystems, their quantity and rate of increase places a burden on our partners and users attempting to efficiently consume and utilize this information. The Template for Assessing Climate Change Impacts and Management Options (TACCIMO) was created to address this need for a standardized, credible, and concise science delivery tool relevant to forest planning and management [5]. As a collaborative effort of the Eastern and Western Forest Environmental Threat Assessment Centers (EFETAC and WFETAC), along with the Southern (R8) and Pacific Southwest (R5) Regional Forest Planning units of the USDA Forest Service, TACCIMO provides an interactive resource for consuming the combination of climate change science and related peer-reviewed literature and land management planning options.

The work highlighted here focuses on TACCIMO’s Geographic Information Systems (GIS) Viewer, a web-based geospatial user interface that allows users to explore climate change science model visualizations within the conterminous United States. The GIS Viewer addresses the demand from research and industry partners for a tool that combines geospatial visualization and summarization.

In this current release, the 3.X version aims to move away from custom-built point-solutions specific to individual needs towards a widely usable integration of geospatial applications. This move from a data repository toward support and guidance includes integration of multiple external hosted data sources into the TACCIMO framework, particularly the TACCIMO GIS Viewer. One of the main changes to the updated TACCIMO GIS Viewer is integrating externally hosted data while minimizing development costs by minimizing changes to existing source code. We also present examples of advanced geospatial analysis methods that leverage this integrated approach and support our goal of providing support, guidance, and unique data interpretations for new and existing partnerships.

This paper is structured as follows: Section two, related work; Section three, architecture and design; Section four, data integration and spatial analysis paradigms; Section five, case studies; and Section six, conclusion and future work.

II. RELATED WORK

There are a wide range of frameworks and techniques for integrating applications, some of which are specific to their domain or a particular technology, while others are generalizable. We present a few as follows.

Service-oriented architectures (SOA) and their associated web services facilitate communication between separate, independent components and specify the protocols governing these communications; this in-turn allows for integration and interoperability of stand-alone applications over the web. Many frameworks and techniques build upon these concepts. SOAP [6] and representational state transfer (REST) [7] are two widely-known and widely-utilized technologies towards these ends, which are in turn used in other techniques and architectures. We too adapt REST in our communications.

Creating software federations is another approach to the ever persistent software integration problem. Some

ISBN: 978-1-61208-326-1

approaches look at Commercial-Off-the-Shelf components [8] for design and integration. Aspect-oriented programming or AOP [9][10] is used in several federation building concepts. AOP separates structures which cross-cut programming abstractions. One approach [11] uses AOP to refactor middleware systems, while another [12] uses AOP as the middleware itself. In techniques applicable to many domains, AOP is used as an exchange for manipulating and processing data between federation components [13]. Our work uses AOP concepts without use of an AOP library.

Tuple-spaces [14][15][16], a generative communication technique coordinating indirect communication between programs, is used in conjunction with several techniques. XMLSpaces [17] is an extensible markup language (XML) focused middleware solution extending IBM's TSpaces implementation [18]. The communication aspects approach [19] uses AOP to uncouple communication and computation, combined with tuple-spaces for communication between aspects. We adapt the communication aspects approach for use in an environment sans AOP.

There are approaches specific to GIS and geospatial analysis domains, such as the webGIS model [20], which uses web 2.0 services to integrate heterogeneous geospatial/geographic databases. Sahina [21] and Aydin [22] amongst others look at GIS-centric usage of SOA and web-services for sharing data between GIS applications. Our approach focuses on interoperability without an overarching architecture.

III. Architecture and System Design

The architecture and system designs descriptions for the TACCIMO GIS Viewer and the associated independent geospatial applications present a high-level view into their underlying structures.

A. TACCIMO Architecture and Design

TACCIMO's overall system architecture for the hosting and technology stack utilizes a multi-tiered system. Content for peer-reviewed literature and planning data are accessed through relational database queries managed by Microsoft SQL Server. Internally hosted geospatial data is managed using Esri ArcGIS Server 10. Development for the TACCIMO web application hosted on the web server was done in Visual Studio. The front-end for the TACCIMO GIS Viewer was built in Flash Builder 4.6 using the Adobe® Flex 4® SDK. The Esri ArcGIS API for Flex is used to access both internally and externally hosted ArcGIS Server resources. Figure 1 gives an overview of the architecture as used within TACCIMO.

B. Climate Wizard

Climate Wizard [23], created in partnership between The Nature Conservancy, the University of Washington, and the University of Southern Mississippi, is a web-based tool [24] providing analyses and geospatial representations of projected global climate change throughout the world in the form of maps, graphs, and tables. It includes temperature and precipitation measurement changes over a historic and two future time periods, using 16 general circulation models (GCMs) and seven GCM ensembles run against three greenhouse-gas emission scenarios. Climate Wizard provides these averages and change in averages for annual, monthly, and seasonal time frames.

This 'big data' application uses ArcGIS Server REST web-services for geospatial data, with underlying data values served through the Amazon Web Services 'on the cloud.' While Climate Wizard provides global climate change predictions, we focused on those for the conterminous United States.

C. California MC1

MAPSS-CENTURY 1 (MC1), created in partnership between Oregon State, Colorado State, and the USDA Forest Service Pacific Northwest Region (R6), provides a dynamic global vegetation model (DGVM) based on the interaction between the MAPSS biogeography model, the CENTURY biogeochemistry model, and a dynamic fire disturbance model [25][26][27]. California MC1 presents the MC1-based predictions for California. The geospatial maps delivered by the California MC1 model are split between the 'MC1 Inputs' for climate change predictions and 'MC1 Outputs' for vegetation and other associated change predictions.

The parameters used for MC1 Inputs are similar to those used in Climate Wizard: two general circulation models run against two greenhouse-gas emission scenarios, measuring average temperature and precipitation for annual and seasonal time frames for one historic and four future time periods. For MC1 Outputs, we use the same two GCMs run against the same two emission scenarios for the same historic and future time periods, but only the annual time frame is used. Instead of measuring temperature and precipitation, vegetation and other climate change indicators are used. These geospatial maps are accessed using ArcGIS Server REST web-services.

D. Climate Change Atlas

The Climate Change Atlas, created by the Northern Research Station of the USDA Forest Service, documents the

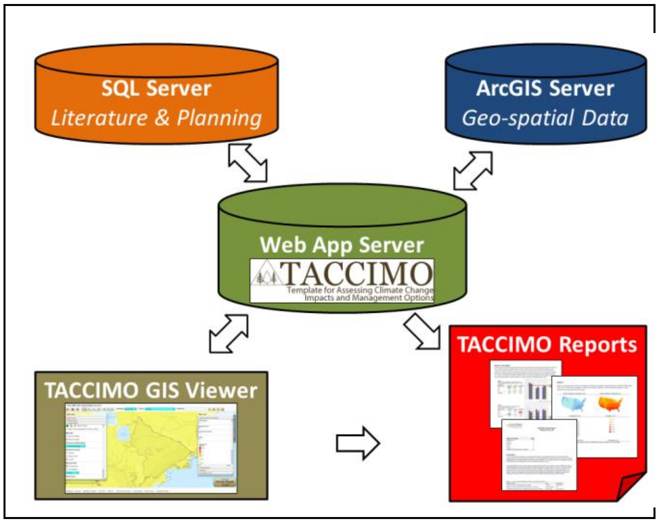


Figure 1. TACCIMO Architecture

ISBN: 978-1-61208-326-1

current and potential habitat distribution shifts for tree and bird species throughout the eastern United States [28][29][30]. The Tree Atlas, a subset of the Climate Change Atlas, delivers the current and possible future distribution for 134 tree species and their underlying environmental predictors using a habitat prediction model (DISTRIB), a colonization likelihood model (SHIFT), and outside factors model (MODFACs).

Habitat suitability is given as importance factor, a measure of relative abundance accounting for tree basal area and the number of stems [31]. The Climate Change Atlas database [32] uses three GCMs and an ensemble average, each run against two greenhouse-gas emission scenarios. Each tree species has a separate model reliability factor. While the original source data is hosted externally, ArcGIS web services are hosted internally.

IV. Software Programming Paradigms

As part of the integration of Climate Wizard, California MC1, and the Climate Change Atlas into the TACCIMO GIS Viewer, we had to address several technical and practical needs: integrating external resources, minimizing changes to existing data, and developing analysis methods using the integrated external data that provide support and guidance to partners and other interested users.

A. Data Integration: Adapted Communication Aspects

The task of moving from internally hosted resources to leveraging data from our collaborative partners was restricted by the limited time and resources of our small development team. The communication aspects technique appeared promising, as it uses aspect-oriented programming and tuple-space modules to connect separate programs into a combined software federation. With Adobe Flex as our programming language for the GIS Viewer front-end, the AOP libraries we explored were not ready for production level use.

AS3Commons AOP, built on AS3Commons Bytecode [33], was an inactive project using older AS3Commons libraries that make it incompatible with our Flex 4 project. It was based on the Loom project [34], commonly cited as the foundation for most AOP endeavors in Flex. Flapper [35], an alternative Flex AOP library based on the Parsley application framework for Flex [36], also had compatibility issues due to development inactivity. The Swiz Framework [37] was another option for AOP within Flex that uses inversion of control and other AOP relevant techniques. Its 2.0 version was slated to include many AOP features we needed, but development of 2.0 was halted and the project donated to Apache Flex in summer 2013 [38], eliminating another option from our list.

Since there were no freely available production-ready AOP libraries for Flex, we looked into bringing the foundational concepts of communication aspects into the GIS Viewer sans AOP. We needed something requiring the effort of one primary developer, ruling out more exhaustive options such as switching languages. In communication aspects, communication and computation are separated using AOP. We achieve a similar separation adapting the ArcGIS Server REST web-services with the observer design pattern [39]; internal communication between components is through the design pattern, while communication between the internal TACCIMO components and data hosted on ArcGIS uses REST.

B. Spatial Analysis: Tree Atlas Summarization

Another thrust of our development changes for the TACCIMO GIS Viewer is developing methods integrating the external data from our partner applications to provide unique data analysis. Here, we explore one of the techniques using a new partner, the Tree Atlas within the Climate Change Atlas.

There are currently distributions for 134 tree species in Tree Atlas; we use the distributions provided across nine climate change model scenarios: three GCMs (Hadley CM3, PCM, and GFDL) and the average of the three GCMs, with all four of these run against two emission scenarios (A1FI and B1), plus the current historical baseline. These combinations make it difficult to synthesize. There is also data in tabular format, but when a user selects a geographic area of interest, these tables often include entries for species not within the area of interest. Inter-species comparisons are limited to two tree species within the selected scenario, even though several dozen trees may exist within the area of interest across the scenarios. The data are there but are not easily consumable for the non-specialist.

Our real-world driven geospatial analysis method is to develop a tool within the TACCIMO GIS Viewer that will, for a user-selected area of interest, create a non-standard visualization comparing all tree species within the area across all nine climate change model scenarios. The original data are arranged by species, yielding 134 data layer tables. We transposed the data matrix into nine data layer tables, allowing for faster data selection and querying through placement in an ArcGIS Server.

Once the user selects the custom area of interest, it is checked for size constraints. The underlying selectable area of interest map is broken down into polygonal cells; if the initial user selected area of interest falls below the constraint threshold, we retrieve the map points bounding each polygon cell in the selected area and run a nearest neighbor search. Those neighboring map points are added to the area of interest. We use this technique to grow the area of interest until it passes the size constraint.

After passing the size constraint check, each climate change model scenario is fetched and queried using REST web-services, bounded by the area of interest. We calculate the adjusted importance value for each species. Since Tree Atlas returns the importance value for each species respective to its entire range of habitability, we average that value over the selected area of interest. An adjusted importance factor of 100 would indicate a monotypic forest stand within the selected area of interest for the climate change scenario. After all scenarios are completed, each species is summed across all climate change scenarios, and those with zero occurrences across all scenarios are pruned. Finally, the species for the current historic climate change scenario are sorted by relative importance from highest to

ISBN: 978-1-61208-326-1

lowest; the eight other scenarios then use the same ordering for their tree species.

V. CASE STUDIES

We show several case studies using two national forests (Francis Marion National Forest and Yosemite National Forest) of interest to our research and industry partners, highlighting the integration of the Climate Wizard, California MC1, and Climate Change Atlas external data sources into the TACCIMO GIS Viewer, plus the utility of our advanced geospatial analysis technique created for the Climate Change Tree Atlas.

A. TACCIMO GIS Viewer Usage Overview

The TACCIMO GIS Viewer is organized around customized themes. There are themes for various case studies requested by our collaborators and partners, in addition to the new themes for California MC1 and the Climate Change Atlas. Each theme provides parameters for users to select which geospatial maps to display in the visualization, along with identification and other spatial analysis tools. Users can also generate a climate change report linking the visualization with peer-reviewed literature and planning language.

B. Climate Wizard

The Climate Wizard integration is available across all themes in the GIS Viewer, as the climate change scenarios it provides are not specific to a particular region or case study. Our parameterization yields 9,384 future climate change scenario permutations: 16 GCMs and seven GCM ensembles, run against three greenhouse-gas emission scenarios (A2, A1B, B1), for 17 time frames (12 months, four seasons, and annual) in three temporal dimensions (future time periods), for the average and changes in average temperature and precipitation. The current historic time period gives 68 permutations, as the GCM and greenhouse-gas emission scenarios do not apply in its case.

Figure 2 shows the change in average annual temperature using the ensemble average GCM and the low (B1) emission scenario projected for the mid-21st century time period at Yosemite, and Figure 6 shows the same climate change projection for Francis Marion. Both show an increase in the mean temperature of approximately four degrees Fahrenheit.

C. California MC1: Yosemite

The California MC1 integration is available through the California MC1 theme. There is a further split of the geospatial visualization layering: MC1 Inputs for climate change predictions and MC1 Outputs for vegetation and other associated change predictions.

MC1 Inputs parameterization yields 160 future climate change scenario permutations: two GCMs (GFDL and PCM) run against two greenhouse-gas emission scenarios (A2 and B1), for five time frames (four seasons and annual) in four temporal dimensions, for the average temperature and precipitation. The current historic time period gives 40 permutations, as the GCM and greenhouse-gas emission scenarios do not apply in its case.

The MC1 Inputs parameterization yields 320 future climate change scenario permutations: the two GCMs run against two greenhouse-gas emission scenarios for five time frames in four temporal dimensions are all identical to those in MC1 Inputs. They measure four climate change indicators: biomass consumed by fire, stream flow depth, maximum snow precipitation, and vegetation habitat classifications. The current historic time period gives 80 permutations.

Here, we look at Yosemite National Forest in California. Figure 3 shows the MC1 Input projected change in average annual temperature using the GFDL GCM run against the B1 low emissions scenario projected for the mid-21st century. This is matched with the MC1 Output projected change in maximum snow precipitation for the same climate change scenario in Figure 4 and the MC1 Output historic maximum snow precipitation in Figure 5. Under this low emissions scenario, users noticed the projected climate impact is a decrease over time in snow along the western slopes of the Sierra Nevada mountain chain in Yosemite National Forest, correlating with the projected temperature increase for the same area in Figure 3.

Having these external climate change science tools gathered in a shared user interface allows decision makers to visually note the projected changes and impacts, and make comparisons. Users can now simultaneously view changes and impacts, gaining insight into interactions between indicators across a large array of climate change scenario configurations.

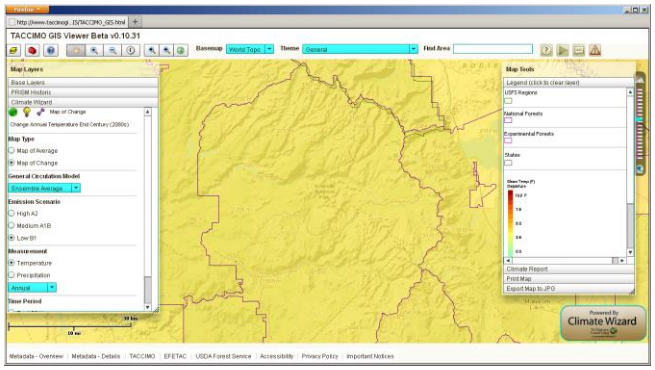


Figure 2. Yosemite projected change in average annual temperature using GCM average ensemble and low B1 emissions in Climate Wizard

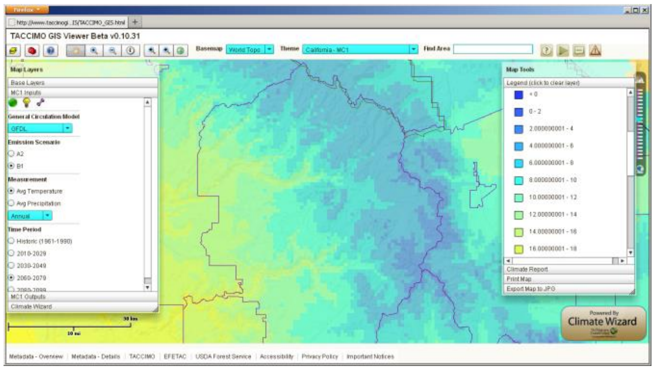


Figure 3. Yosemite projected change in average annual temperature using GFDL and low B1 emissions in California MC1 Inputs

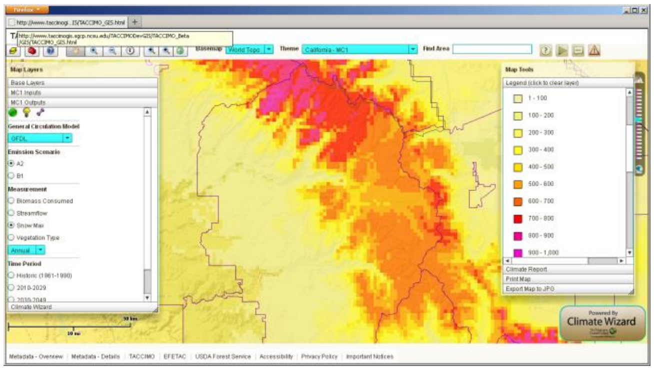

Basemap: World Topo
Theme: California - MC1
General Circulation Model: GFDL
Emission Scenario: ● A2 ○ B1
Measurement: ○ Biomass Consumed ○ Streamflow ● Snow Max ○ Vegetation Type
Annual
Time Period: ○ Historic (1961-1990) ○ 2010-2029 ○ 2030-2049
http://www.taccimogi...IS/TACCIMO_GIS.html
http://www.taccimogis.sgcp.ncsu.edu/TACCIMODevGIS/TACCIMO_Beta/GIS/TACCIMO_GIS.html
Find Area
Map Layers
Base Layers
MC1 Inputs
MC1 Outputs
Climate Wizard
Map Tools
Legend (click to clear layer)
1 - 100
100 - 200
200 - 300
300 - 400
400 - 500
500 - 600
600 - 700
700 - 800
800 - 900
900 - 1,000
Climate Report
Print Map
Export Map to JPG
30 km
10 mi
Powered By Climate Wizard
Metadata - Overview | Metadata - Details | TACCIMO | EFETAC | USDA Forest Service | Accessibility | Privacy Policy | Important Notices

Figure 4. Yosemite projected change in maximum annual snow precipitation using GFDL and low B1 emissions in California MC1 Ouputs

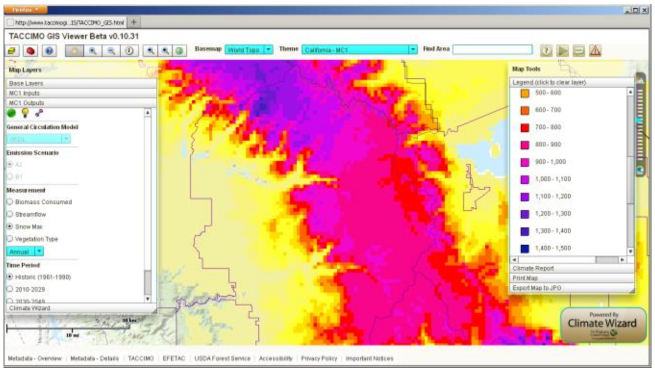

TACCIMO GIS Viewer Beta v0.10.31
Basemap: World Topo
Theme: California - MC1
General Circulation Model: GFDL
Emission Scenario: ● A2 ○ B1
Measurement: ○ Biomass Consumed ○ Streamflow ● Snow Max ○ Vegetation Type
Annual
Time Period: ● Historic (1961-1990) ○ 2010-2029 ○ 2030-2049

Figure 5. Yosemite current historic maximum annual snow precipitation in California MC1 Outputs

D. Climate Change Tree Atlas: Francis Marion

The Climate Change Tree Atlas integration is available using the Climate Change Atlas theme. Along with the visualization of suitable habitats for each of the 134 tree species, there is a tool for selecting an area of interest and generating a non-standard visualization summarizing the relative abundance of each tree species.

Our Tree Atlas parameterization yields three scenarios for each tree species: current historic, average GCM for the A1FI (high emissions) scenario, and average GCM for the B1 (low emissions) scenario. We do not use any of the temporal dimensions for the GCMs. The Climate Change Tree Atlas summarization tool allows the user to select an area of interest, and then calculates and generates a chart showing the relative importance of each tree species. It uses all nine climate change scenarios available through Tree Atlas: the Hadley CM3, PCM, and GFDL GCMs, plus the average GCM run against the A1FI and B1 emissions scenarios, along with the current historic baseline scenario.

Here, we look at Francis Marion National Forest along the coast of South Carolina. We used the GIS Viewer to explore scenarios using *Pinus taeda* or loblolly pine, the dominant tree species for that area [40]. Comparing the GCM average run against the B1 low greenhouse-gas emissions scenario in Figure 7 with the historic in Figure 8, users noted a decrease in relative habitability dominance for the tree species. Users then asked several questions: “is loblolly pine still the dominant species with regards to habitat in future scenarios, how much does it decrease relative to the other species, and does it decrease enough so that another surpasses it with respect to habitat suitability?”

The charts in Figure 9 and 10 generated by the Tree Atlas advanced geospatial analysis summarization tool allowed users to quickly look further into these questions without having to manually search each of the 130+ other tree species across the various climate change scenarios. In Figure 10, users hovered their mouse over the data point for the GCM average run against B1 of loblolly pine, showing its adjusted importance value at around 15.53. Users noted the spike for *Pinus elliottii* or slash pine in the same GCM average run against B1, with its adjusted importance value at approximately 10.98. Forest, land management, and other experts then use this information towards further investigation.

VI. CONCLUSION AND FUTURE WORK

The current development stage of the TACCIMO GIS Viewer allows for integrating externally hosted geospatial data sources with peer-reviewed literature and planning language. Internally, this integration was managed by a combination of aspect-oriented programming-like observer patterns, with external communications centered on ArcGIS REST web-services. These integrations allowed policy developers and land managers to conveniently investigate multiple climate change science resources using interactive visualizations for a wide range of permutations parameterized using GCMs, greenhouse-gas emission scenarios, temporal dimensions, time frames, and climate change measurement indicators. We then used the integrated data to create a new, non-standard, advanced geospatial analysis. These techniques aid our roles in providing support and guidance to our users, who in turn use this to assess, manage, and monitor forest resources. Small numbers of users are responsible for a large amount of land, meaning large land management impact.

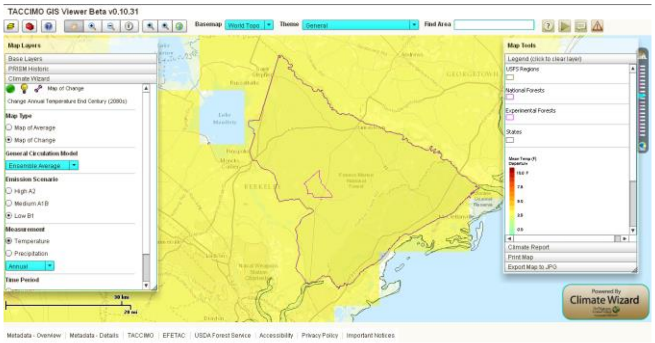

TACCIMO GIS Viewer Beta v0.10.31
Basemap: World Topo
Theme: General
Map of Change
Change Annual Temperature End Century (2080s)
Map Type: ○ Map of Average ● Map of Change
General Circulation Model: Ensemble Average
Emission Scenario: ○ High A2 ○ Medium A1B ● Low B1
Measurement: ● Temperature ○ Precipitation
Annual
Time Period

Figure 6. Francis Marion projected change in average annual temperature using GCM average ensemble and low B1 emissions in Climate Wizard

ISBN: 978-1-61208-326-1

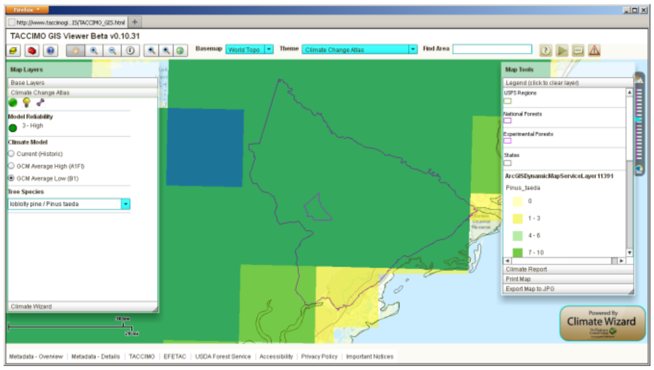

TACCIMO GIS Viewer Beta v0.10.31
Basemap World Topo
Theme Climate Change Atlas
Find Area
Map Layers
Base Layers
Climate Change Atlas
Model Reliability
3 - High
Climate Model
○ Current (Historic)
○ GCM Average High (A1FI)
● GCM Average Low (B1)
Tree Species
loblolly pine / Pinus taeda
Climate Wizard
Map Tools
Legend (click to clear layer)
USFS Regions
National Forests
Experimental Forests
States
ArcGISDynamicMapServiceLayer11391
Pinus_taeda
0
1 - 3
4 - 6
7 - 10
Climate Report
Print Map
Export Map to JPG
Powered By Climate Wizard
Metadata - Overview | Metadata - Details | TACCIMO | EFETAC | USDA Forest Service | Accessibility | Privacy Policy | Important Notices

Figure 7. Francis Marion projected change in relative importance of loblolly pine using GCM average for B1 emissions in Tree Atlas

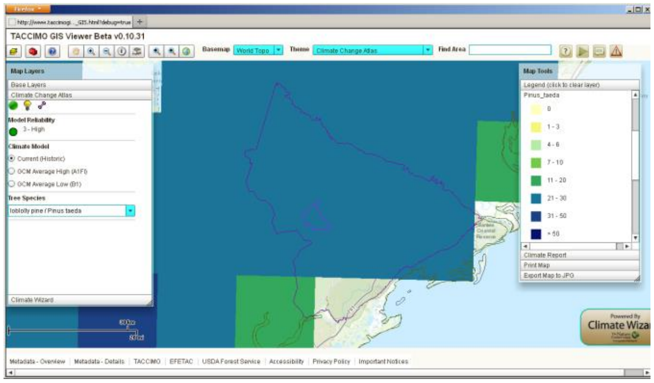

TACCIMO GIS Viewer Beta v0.10.31
Basemap World Topo
Theme Climate Change Atlas
Find Area
Map Layers
Base Layers
Climate Change Atlas
Model Reliability
3 - High
Climate Model
● Current (Historic)
○ GCM Average High (A1FI)
○ GCM Average Low (B1)
Tree Species
loblolly pine / Pinus taeda
Climate Wizard
Map Tools
Legend (click to clear layer)
Pinus_taeda
0
1 - 3
4 - 6
7 - 10
11 - 20
21 - 30
31 - 50
> 50
Climate Report
Print Map
Export Map to JPG
Powered By Climate Wizard
Metadata - Overview | Metadata - Details | TACCIMO | EFETAC | USDA Forest Service | Accessibility | Privacy Policy | Important Notices

Figure 8. Francis Marion current historic relative importance of loblolly pine in Tree Atlas

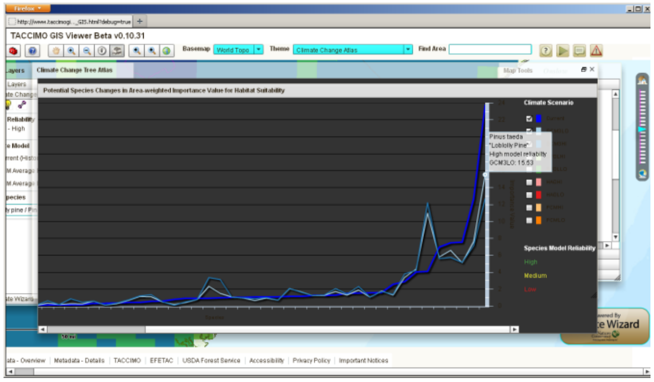

TACCIMO GIS Viewer Beta v0.10.31
Basemap World Topo
Theme Climate Change Atlas
Find Area
Climate Change Tree Atlas
Map Tools
Potential Species Changes in Area-weighted Importance Value for Habitat Suitability
Climate Scenario
Pinus taeda
"Loblolly Pine"
High model reliability
GCM3LO: 15.53
Species Model Reliability
High
Medium
Low
Metadata - Overview | Metadata - Details | TACCIMO | EFETAC | USDA Forest Service | Accessibility | Privacy Policy | Important Notices

Figure 9. Francis Marion advanced geospatial analysis chart highlighting relative importance of loblolly pine in GCM average for B1 emissions

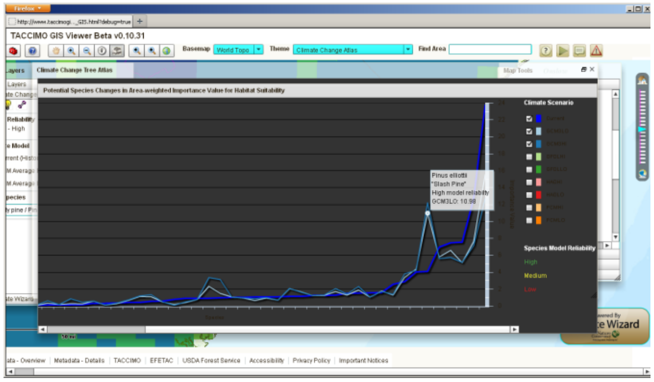

TACCIMO GIS Viewer Beta v0.10.31
Basemap World Topo
Theme Climate Change Atlas
Find Area
Climate Change Tree Atlas
Map Tools
Potential Species Changes in Area-weighted Importance Value for Habitat Suitability
Climate Scenario
Pinus elliottii
"Slash Pine"
High model reliability
GCM3LO: 10.98
Species Model Reliability
High
Medium
Low
Metadata - Overview | Metadata - Details | TACCIMO | EFETAC | USDA Forest Service | Accessibility | Privacy Policy | Important Notices

Figure 10. Francis Marion advanced geospatial analysis chart highlighting relative importance of slash pine in GCM average for B1 emissions

Future work can be explored on several fronts. From the external data integration front, additional content sources, such as the Climate Change Bird Atlas, are under consideration. Another area is exploring updates to the climate report generation process. For geospatial tool analysis, the area of interest selection process could be fine-tuned. Domain experts have suggested several ideas, such as adding or removing individual cells, automatically pruning cells based on ecoregion boundaries, and selecting species of interest for summarization charts.

ACKNOWLEDGMENT

The authors give their thanks to Chris Zganjar and Evan Girvetz from The Nature Conservancy, plus George T. Raber from the University of Southern Mississippi, for their help and support in integrating Climate Wizard into the GIS Viewer. We thank John Cobb from NC State University for his extensive and diligent work on the Tree Atlas advanced geospatial analysis tool. We thank John Kim and Charlie Schrader from the USDA Forest Service for their help and support in integrating the California MC1 services. We also thank Louis Iverson, Anantha Prasad, Matt Peters, and Steve Matthews from the USDA Forest Service for their input and feedback during the Climate Change Tree Atlas integration process.

REFERENCES

[1] R. W. Malmsheimer et al., "Forest management solutions for mitigating climate change in the U.S.," Journal of Forestry 106, 2008, pp. 115-171.

[2] Climate Change Resource Center: Tools for Land Managers, www.fs.fed.us/ccrc/tools/, accessed: 2014-01-06.

[3] C. Swanston and M. Janowiak, eds., "Forest adaptation resources: Climate change tools and approaches for land managers," US Department of Agriculture, Forest Service, Northern Research Station, Newtown Square, PA, Gen. Tech. Rep. NRS-87, 2012.

[4] D. L. Peterson et al, "Responding to climate change in national forests: a guidebook for developing adaptation options," US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR, Gen. Tech. Rep. PNW-GTR-855, 2011.

[5] L. Jennings, E. Treasure, J. Moore Myers, and S. McNulty, "The Template for Assessing Climate Change Impacts and Management Options (TACCIMO): Science at your fingertips," Poster presented at 2012 Fall Meeting, American Geophysical Union (AGU), San Francisco, CA, Dec. 2012.

[6] M. Gudgin et al, "SOAP Version 1.2 Part 1: Messaging framework (Second Edition)", W3C Recommendation, April 27, 2007, www.w3.org/TR/soap12/, accessed: 2014-01-06.

[7] R. T. Fielding and R. N. Taylor, "Principled Design of the Modern Web Architecture," ACM Transactions on Internet Technology 2 (2), May 2002, pp. 115-150.

ISBN: 978-1-61208-326-1

[8] J. Estublier, H. Verjus, and P.-Y. Cunin, “Designing and building software federations,” Proc. 1st Conference on Component Based Software Engineering (CBSE-EUROMICRO 2001), Warsaw, Poland, Sept. 2001, pp. 121-129.

[9] G. Kiczales et al., “Aspect-oriented programming,” Proc. 11th European Conference on Object-Oriented Programming (ECOOP 1997), pp. 220-242, doi: 10.1007/BFb0053381

[10] C. V. Lopes, “AOP: a historical perspective,” Proc. Aspect-Oriented Software Development, R.Filman et al (eds.), 2004.

[11] C. Zhang and H.-A. Jacobsen, “Quantifying aspects in middleware platforms,” Proc. 2nd International Conference on Aspect Oreinted Software Development (AOSD 2003), Boston, MA, Mar. 2003, pp. 130-139.

[12] A. Colyer and A. Clement, “Large-scale AOSD for middleware,” Proc. 3rd International Conference on Aspect-Oriented Software Development (AOSD 2004), Lancaster, UK, 2004, pp. 56-65.

[13] D. Stotts, K. L. Lee, and I. Rusyn, “Supporting computational systems science: genomic analysis tool federations using aspects and AOP,” Proc. 4th ISBRA, Atlanta, GA, May 2008, pp. 457-469, doi: 10.1007/978-3-540-79450-9_43.

[14] D. Gelernter, “Generative communication in Linda,” ACM Transactions on Programming Languages and Systems 7(1), 1985, pp. 80-112.

[15] S. Ahuja, N. Carriero, and D. Gelernter, “Linda and friends,” Computer 19 (8), 1986, pp. 26-34.

[16] N. Carriero, D. Gelernter, “Linda in context,” Communications of the ACM 32 (4), 1989, pp. 444-458, doi: 10.1145/63334.63337.

[17] R. Tolksdorf, F. Liebsch, and D. M. Nguyen, “XMLSpaces.NET: and extensible tuplespace as XML middleare,” Proc. 2nd International Workshop on .NET Technologies. 2004.

[18] P. Wyckoff, S. W. McLaughry, T. J. Lehman, and D. A. Ford, “TPSaces,” IBM Systems Journal 37 (3), 1998, pp.454.

[19] K. L. Lee and D. Stotts, "Composition of bioinformatics model federations using communication aspects," Proc. IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, PA, Oct. 2012, doi:10.1109/BIBM.2012.6392621

[20] M. Pascaul, E. Alves, H. Roig, T. de Almeida, G. S. de Franca, and M. Holanda, “An architecture for geographic information systems on the web – webGIS,” Proc. 4th International Conference on Advanced Geograpic Information Systems, Applications, and Services (GEOProcessing 2012), Valencia, Spain, 2012, pp. 209-214.

[21] K. Sahina and M. U. Gumusayb,“Service oriented architecture (SOA) based web services for geographic information systems,” Proc. 21st ISPRS Congress, Beijing, China, pp. 625-630.

[22] G. Aydin, “Service Oriented Architecture for Geographic Information Systems Supporting Real Time Data Grids”, Ph. D. Dissertation, Department of Computer Science, Indiana University, 2007.

[23] E. H. Girvetz, C. Zganjar, G. T. Raber, E. P. Maurer, P. Kareiva, and J. J. Lawler, “Applied climate-change analysis: The Climate Wizard tool,” PLoS ONE 4 (12), Anna Traveset, Ed., Dec. 2009, doi:10.1371/journal.pone.0008320.

[24] ClimateWizard, www.ClimateWizard.org, accessed: 2014-01-06.

[25] J. M. Lenihan, C. Daly, D. Bachelet, and R. P. Neilson, “Simulating broad-scale fire severity in a dynamic global vegetation model,” Northwest Science 72 (2), 1998, pp. 91-103.

[26] C. Daly, D. Bachelet, J. M. Lenihan, R. P. Neilson, W. Parton, and D. Ojima, “Dynamic simulation of tree-grass interactions for global change studies,” Ecological Application 10 (2), 2000, pp. 449-469.

[27] D. Bachelet, J. M. Lenihan, C. Daly, R. P. Neilson, D. S. Ojima, and W. J. Parton, “MC1. A dynamic vegetation model for estimating the distribution of vegetation and associated ecosystem fluxes of carbon, nutrients and water,” General Technial Report PNW-GTR-508, Corvalis, OR, USDA Forest Service, 2001.

[28] A. M. Prasad, L. R. Iverson, S. Matthews, and M. Peters, “A Climate Change Atlas for 134 Forest Tree Species of the Eastern United States [database],” Northern Research Station, USDA Forest Service, Delaware, Ohio, 2007-ongoing, www.nrs.fs.fed.us/atlas/tree, accessed: 2014-01-06.

[29] L. R. Iverson, A. M. Prasad, S. N. Matthews, and M. Peters, “Estimating potential habitat for 134 eastern US tree species under six climate scenarios,” Forest Ecology and Management 254, 2008, pp. 390-406.

[30] A. M. Prasad, L. R. Iverson, S. Matthew, and M. Peters, “Atlases of tree and bird species habitats for current and future climates,” Ecological Restoration 27, 2009, pp 260-263.

[31] S. G. McNulty et al., “Application of linked regional scale growth, biogeography, and economic models for southeastern United States pine forests,” World Resource Review 12 (2), pp 298-320, 2000.

[32] Climate Change Atlas, www.fs.fed.us/nrs/datlas/, accessed: 2014-01-06.

[33] AS3Commons Bytecode, www.as3commons.org/as3-commons-bytecode/, accessed: 2014-01-06.

[34] Loom, code.google.com/p/loom-as3/, accessed: 2014-01-06.

[35] Flapper, code.google.com/p/flapper-as3/, accessed: 2014-01-06.

[36] Parsley Framework, www.spicefactory.org/parsley/, accessed: 2014-01-06.

[37] Swiz Framework, swizframework.jira.com/wiki/, accessed: 2014-01-06.

[38] Apache Flex ready for Swiz donation, groups.google.com/d/topic/swiz-framework/XUMHxelq_-4, accessed: 2014-01-06.

[39] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, “Design patterns: elements of reusable object-oriented software,” Addison-Wesley, 1995,. ISBN 0-201-63361-2.

[40] Francis Marion Land and Resource Management Plan, www.fs.usda.gov/goto/scnfs/fmplan, accessed: 2014-01-06.

Geographical Information Systems Participating into the Pervasive Computing

Nuhcan Akçit
Department of Geodetic and Geographic Information Technologies
Middle East Technical University
Ankara, Turkey
nuhcan@metu.edu.tr

Emrah Tomur
Department of Information Systems
Middle East Technical University
Ankara, Turkey
etomur@metu.edu.tr

Mahmut Onur Karslıoğlu
Department of Civil Engineering,
Geomatics Engineering Division,
Department of Geodetic and Geographic Information Technologies
Middle East Technical University
Ankara,Turkey
karsliog@metu.edu.tr

***Abstract*—The popular term "pervasive computing" describes the concept of computers being everywhere. The Geographic Information System (GIS) technology began with completely centralized systems after the influence of the Internet; the GIS then changed into decentralized systems. In the last few years, the GISs have been included in many areas of pervasive computing, such as the Internet of Things and sensor networks. Changing and developing states of the GIS such as Internet, mobile, web GIS, ubiquitous GIS, the Internet of Things and sensor network are the different elements of the evolution. From the beginning of the evolution until today, pervasive computing is described in this study. In addition to GIS as a part of pervasive computing technology, the activity of Open Geospatial Consortium (OGC), which works on geospatial and location standards in these technologies, is also discussed. This paper investigates the evolution of GISs from centralized systems to a distributed system, summarizes the steps and use areas of the GISs to participation in pervasive computing, and compares different GISs concepts in application areas. In addition, future trends and expectations of the GISs are included.**

Keywords-GIS; pervasive computing; WEB GIS; mobile GIS; Internet of Things

I. INTRODUCTION

The Geographic Information System (GIS), a system designed to store, capture, analyze, manipulate, and manage all types of geographic data, is an interdisciplinary area that combines cartography, analysis, statistics, and computer technology [1].

In recent years, ubiquitous and pervasive computing domains have increased considerably. They are a part of the daily life because of the technological evolution and new applications. In this paper, the goal is to provide information about geographic information science, a system that has developed considerably and is involved in our lives more than ever. The fundamental areas of geographic information science are remote sensing, surveying, photogrammetry, and analyzing spatial data.

The focus of this paper is GISs, which are becoming part of the ubiquitous computer domain. In this paper, we provide a brief information about the GIS and its development from the 1990s to 2010. We mainly focus on after 90's because distributed and Internet GIS [2] started to be formed in this period, and our ubiquitous domain is more interested in development after 90's. We also discuss future trends in this area. The GISs are huge; therefore, we focus on the development and evolution of the ubiquitous and pervasive computing domains. Thus, elements of the development of the GISs are beyond the scope of this paper.

The GIS is not familiar to the public. We first provide general information about the GIS and application areas. Then, we discuss the evolution of the GIS from the 1990s to today [2]. In the last few years, the GISs have become a larger part of the ubiquitous domain [7]. This study is a brief summary based on books, articles, and other sources. We mainly investigate information about the GIS, development stages, and evolution steps until today.

This paper is structured as follows: We discuss GISs in Section 2. The effects of the Internet on GISs are analyzed in Section 3. Data Centralized-distributed GISs are considered in the Section 4. Decentralized distributed GISs are considered in Section 5 (location-based services, web GIS, mobile GIS, the open geospatial consortium, ubiquitous GIS sensor network, and the Internet of Things).

II. THE GEOGRAPHIC INFORMATION SYSTEMS

Generally, GISs are systems that "integrate hardware, software, and data for capturing, managing, analyzing, and displaying the different forms of geographically referenced information" [3]. All types of geographically references data can be organized by The Geographic Information Systems. Today, the major components of GISs are people, hardware, software, data procedures, and the network [3]. A GIS allows us to understand all types of reports, data, and visualize, interpret, and find relationships between them. A GIS helps people answer questions and solve problems by looking for data in a way that is quickly understood and easily shared. GIS technology is used by any kind of work which can be integrated easily with GIS. Spatial analysis is mainly used in many areas; thus, with the help of the GIS, analyzing data and finding results in different domains is easy, a key element of the GIS.

General benefits of using the Geographic Information Systems include saving money, increasing efficiency, aiding in making better decisions, improving communication, keeping records better and deleting the managing data geographically or the relations [4]. Different types of data management and different types of software deal with different types of data. However, these details are beyond the scope of this paper. In many open source and commercial programs, the usage is based on the work someone wants to do exactly, which means the selection changes with the context of the work. Only basically say that the data types include different raster, vector and hybrid format[3]. The data sources include images, digitized or scanned maps, GPS data, and data transferred from other places with the required program.

What is unclear about the GISs are whether; systems are inter-disciplinary or a separate discipline. According to some, the GISs are inter-disciplinary; others believe since the system has been developed, it is a separate discipline. Since this issue does not affect the capability or development of the GIS, this issue is not important in this paper. This paper focuses on GIS capabilities and how the GISs have affected daily life since being developed in the 1990s.

Now looking at some little information on how GIS is developed at first. The GIS was first developed in Canada in the mid-1960s to identify land resources, where they were, and how they were used. In the late 1970s, Harvard University's Laboratory for Computer Graphics and Spatial Analysis developed a general purpose GIS [1][5]. In the early 1980s, the price of computing hardware decreased, which helped sustain the software industry and led to the development of cost-effective applications. This is making the more powerful and efficient standalone systems which have been introduced in the 1980’s. During the 1990s, map analysis and modeling start to increase this software programs. After the 1990s, the Internet grew rapidly, which led to new concepts: distributed GIS and Internet GIS [2].

III. Effect of the Internet on the Geographic Information Systems

The Internet grew rapidly in the 1990s, and has become part of society. Ubiquitous access to the Internet makes it more powerful for people to process, exchange, and access information. The Internet also changed GIS data access, manipulating, and sharing. The Internet led to an important change and improved GIS in terms of spatial analysis. The Internet affected the GIS in three areas: data access, spatial information distribution, and processing [2]. The Internet provides analytical results for a much wider audience than the traditional GIS. People start to search online and use free search and query analysis for spatial objects without purchasing expensive GIS software. In the processing field, the Internet enhances the reusability and accessibility of the analysis tools and dynamic download and upload, which also helps interactive work on data, which is discussed later in this current study.

Accessing data over the Internet was the first step for the Internet GIS [2]. Accessing data online allows people who have stand-alone GIS on their local machines to transfer data over the Internet. This method is suitable for accessing data quickly and easily but harder for the analyzing data independent from the location, which is the next important step. Online GIS processing continues to improve to increase the analytical abilities of the Internet.

IV. data Centralized- Distributed GIS

The GIS influenced the advancement of information technology. The development of the GIS resembles computer development, which evolved from a mainframe GIS to a desktop GIS [2] and then distributed GIS, which includes Internet GIS and mobile GIS.

Distributed GIS has subtypes which are Local Area Network (LAN) based GIS, Internet based GIS and mobile based GIS. In this paper we are interested with the distributed GIS refers to GIS programs working on the internet (Internet GIS) or wireless network environments (mobile GIS)[2].Distributed GIS is mainly based on improving the Internet and wireless technologies based on the rapid expansion of low-cost bandwidth [2] over the Internet, and web-enabled desktop and mobile devices.

The distributed GISs differentiate the client-server LAN-based traditional method [2], which does not require the installation of a GIS program on the user’s desktop. This method relies on the access to analytical tools and data anywhere through Internet access or wireless coverage. The client is any device that can be connected to the Internet or may be connected to a wireless capable device to access the data or tool. There are two concepts in distributed GIS: Internet GIS [2] (can be wired or wireless also) and mobile GIS (wireless). This is important because depending on whether communications are wired or wireless, the applications may change. Although these issues are very important for developing a GIS, the interoperability of the data is very important.

V. Decentralized Distributed GIS

In distributed GISs, many names are used: online GIS, distributed geographic information (DGI), web-based GIS, or web GIS [2][6]. Internet GIS and web GIS sound as if they are the same, but they are completely different from each other. Internet GIS involves using the Internet to exchange data, perform analysis, and present results; however, web GIS primarily uses the World Wide Web (WWW). Internet GIS and web-based GIS use the client server as a model [2][7]. Web GIS uses web browsers, which are a major part of the Internet, but Internet GIS is broader and, it uses more than browsers.

Distributed GIS mainly uses the client server as a model [2][7]. The server itself performs the job and sends the result to the client or sends the data and analysis tool to the client to calculate the result. The connections and the communication are based on the standards of the client server or Internet standards (which are not discussed in this section because they are the same as the general use of Internet standards and the client-server architecture). The thick and thin client model can be also seen as distributed GIS [2].

Distributed GIS are hosted on several hosts connected by a communication network. One subclass of such distributed GIS is Internet GIS. WebGIS is a subclass of Internet GIS which is significantly more usage last years based on the WWW (World Wide Web), and its add-ons to provide interactivity between the user and the distributed GIS program. To enable interactivity, HTML (Hyper Text Markup Language), XHTML (Extensible Hyper Text Markup Language), WAP (Wireless Application Protocol), and vector-based GIS can be used [2]. Users can then manipulate the data and maps interactively. Users also can perform GIS functions such as rendering, spatial queries, and spatial analysis using a web browser or other Internet-based applications.

Distributed GISs take advantage of the Internet as a distributed system [2], which means GIS data and the analysis tool can be on different computers over the Internet. The data can be in different places at a company or institution and, it can be accessed, combined, or analyzed with help of the Internet or intranet. In addition, public agencies can provide data for users or other providers. Individuals can search and download data or tools over the Internet. Choosing different applications for different types by the same user is possible. The main advantage of distributed GIS [2] is being connected dynamically to data sources. Real-time information can be used in a real-time connection such as a real-time satellite image, emergency response information, and traffic movement.

Another important issue is the accessibility of the GIS from different platforms. It must not be limited to one type of a system. As long as people have connections to the Internet, individuals can access and use distributed GIS. Cross-platform and inter-operable GIS tools [2] are important, so everybody can access or operate the tools independently of the system they use. In addition, interoperability means individuals can access the GIS from various devices. The Open Geodata Interoperability specification [2] and geography markup language (GML) [8] by the Open GIS Consortium (OGC) set standards and rules for interoperability [2][8]. (There are also many different standards for other concepts that are created by the OGC during 2000s, which will be discussed in detail later.)

Distributed GISs are required for many purposes. One is the uniqueness of the data over the Internet. For example, a road map is more usable by tourists if there are points of interest on the map (hotels, parks, restaurants). This example shows the importance of using geographic data increasing with the inter-operability and cross-platform issues [8], people rarely focus on this issue for actual implementation in the beginning. These needs must be achieved in cross-platform programs when a system is designed in concept of the GIS. It must be well defined to make the program truly inter-operable and cross-platform. This requires the community to revise meta-data. This requirement led to research on meta-data for geospatial data and software components in the late 1990s. Integrated meta-data is one of the successful points of distributed GIS.

Distributed GIS is used in many other areas today [2], which means a flexible and dynamic scheme is more usable. Delivering data with traditional GIS has difficulties. Different programs have unique processes and data sets that make life more difficult. Needs are divided into three perspectives: management, user, and implementation perspectives [1][2]. From the management perspective, there are two main reasons why distributed GIS should be more useful. First is the globalization of geographic information series distribution. Data were distributed in papers or paper maps previously, but in recent years, many data sources have become publicly available on the Internet. To provide a GIS community global system, they built data to distribute [6] online in the 1990s. It provided a large scope for conducting scientific research, and GIS users obtain information easily from the community. Second, Internet GIS services are related to decentralization of managing and updating the system [2] [6]. Data gathering techniques such as global positioning systems (GPSs), satellite images, remote sensing data, and other GIS [1][9] application data mean people must deal with huge databases. Since the databases are huge, this creates maintenance and update issues. Internet GIS provides a solution to this situation. Data sets are maintained at the source site instead of a centralized location [2]. Another advantage is the increasing reliability failure of one location. Thus, the entire distributed Geographic Information System does not fail. This increases the efficiency of the system and reduces the cost of maintaining the database.

From the user perspective, there are three main reasons why Internet-based distributed GIS services are important. The first is the huge GIS data sets and processing are not possible. Smaller workstations and a distributed system provide a chance for dynamic processing at servers and encapsulate the results for return to the client. The second is customization of GIS modules. With distributed GIS, individual software modules are updated, which provides more flexible solutions to users. The third reason is the public’s demand for location-based information due to the popularity of the Internet and mobile devices. Popularity began with the GPS, in-car navigation, and wireless access to the Internet. Distributed GIS services have a real potential to bring GIS to the masses [2].

From the implementation perspective, there are three main problems in distributing the system. The first is that Internet-based GIS [2] services lack high-level architecture, which means there are temporary solutions for distributed systems. The second problem is the interoperability issue [8] of data and processes. The third problem is data overload because of the many different sources. Each source may have the same data. This means loss of power and high costs to duplicate the data from different databases. These are solved if the GIS users have the knowledge required to integrate different sections and critical issues.

Distributed GIS has a wide range of applications that can be categorized into four main groups: data sharing and dissemination, simple search and querying, online data processing, and location-based services. In data sharing, the system shares data on the Internet perhaps over a web browser, FTP, or other type of Internet application. Information is disseminated via maps and data processes by specialists. Specialists give users user permission to use these materials. Online processing makes people work on providing inter-operable software development, or they process different parts of the problem and put them together to solve problem with online processing. Another important process is location-based services (LBS) [2][10], which refer to real-time implementation. LBS help people find the best route to their destination. Distributed GIS can meet needs easily. In addition, mobile GIS [2][7], which is described in detail later, can offer mobile users real-time traffic information, landmarks near users, and points of interest.

A. Location-Based Services (LBS)

The development of technology and technological trends has involved more location-based services in daily life. Improvement in distributed GIS also aided in the development of location services.

Technology is rapidly changing, particularly in Geographic Information Systems. GNSS is the global navigation satellite system. One type of GPS (global positioning system) is operated by the United States. However, there is also GLONASS (Globalnaya Navigatsionnaya Sputnikovaya Sistema), which was operated by Russia. Other systems in development include Galileo (the European Union [EU]) and Beidou (China) [9]. There are also GISs that are not GNSS: radio frequency identification and location-sensing technologies with varying accuracy.

Location-based services [2][10] are mainly services or applications that extend spatial information processing to users or GIS capabilities via the Internet or wireless devices. GIS capabilities are based on the networked or distributed concept. Many LBS are used today, including social networking. LBS are targeted at many people, and are inter-operable and cross-platform [8][10]. LBS evolved from online map services and Internet GIS applications. LBS are used for more concentrated lightweight devices [10]; GIS applications are more general. However, LBS definitions and uses overlap with those of the GIS.

Since GIS have interoperability and cross-platform issues, LBS evolved from GIS for more cross-platform and diversity. LBS have more simultaneous and dynamic processes [10] in which data are obtained from the traffic or remote sensed by satellites, which is based on GIS applications and concepts. LBS were developed for more independent use and are useful only for location services. There are different types of LBS; however, since the 2000s mobile phones have been equipped with GPS receivers. Thus, all over the world, this type of LBS is commonly used.

LBS capabilities are limited to finding locations and tracking objects with small sensors [2][10]. Locations are found based on the following questions: Where am I going? How do I get there? Where am I now? LBS work very well in this context. LBS have two modes: pull and push [10]. The push mode services are pushed to the user automatically and are not based on needing the information. In the pull mode, the user requests the data voluntarily to access information.

The object tracking capability of LBS with small sensors is attached to the object that people want to track. However, sensor networks are better at tracking objects, and will be discussed later. The sensor network is better at giving users what they need to track. In the future, ubiquitous computing based on Weiser's visions will help improve [11] LBS.

LBS applications must be designed from the user's perspective in the GIS context. LBS work thanks to the georeferencing system, which may use GPS (or any GNSS system) for global positioning or take the initial location as a georeference and calculate indoors based on geometric models. There are actually two types of modeling: geometric and symbolic modeling. GIS use geometric modeling to calculate LBS. Geometric models have limitations since the public uses more verbal location with expressions of spatial features, location, and spatial relations. Some initiatives are integrating symbolic and geometric modeling. A new concept is the semi-symbolic model that contains a geometric location and a symbolic representation. Context-aware computing is also part of LBS, but the context is the most difficult part [10]. The context affects users' actions, behavior, and information retrieval. The context used by mobile users changes frequently; therefore, the LBS must take the changing issues continuously to provide service to mobile users. Geospatial data are a key concept in LBS [10], which means this topic is related to the GIS. It uses mainly geospatial data and GIS to determine locations. Determining the location is meaningless if it does not refer to a georeference place. If georeferencing is not used, positions are found relative only to other materials; thus, people do not know where they are exactly. In addition, the base layer and additional position information are needed to find locations on the earth.

The model mainly proposed a domain of ubiquitous computing with some exceptions, data modeling conducted in GIS. LBS can be treated as a special type of geographic information services. Research on spatial ontologies has implications for the development of LBS. The uncertainty of geographic information is linked to data processing and modeling in LBS [10].

The ubiquitous use of the GIS [12] affects many concepts in this domain. The next-generation GIS is related more to the ubiquitous domain since information from the mid-2000s reflects technology today. In addition, the OGC has a role as this organization creates regulations regarding the interoperability of LBS in terms of geographic and geospatial aspects.

B. *Web GIS*

The rapid expansion of the Internet has led to new disciplines; Web GIS is one of them. It changes geospatial information by acquiring, publishing, sharing, and visualizing. Web GIS is a milestone in the development of the GIS. In 1993, the first web GIS was introduced, by the Xerox Pablo Alto Research Center (PARC) Web-based map viewer [2][7][11]. It was an experiment for retrieving information on the web that did not directly access data. This viewer provides simple zooming, layer selection, and map projection capabilities. The GIS can be used without local installation.

New web GIS [2][7] applications then emerged: in 1994, the Canada National Atlas Information Service and in 1995, topologically Integrated Geographic Encoding and Referencing (TIGER), a US mapping service. In 1996, MapQuest, a web application that allows people to view maps, look for local businesses, find the optimal routes to destinations, and plan trips, was introduced. It is still used today.

Web 2.0 formed at the end of the 1990s and after that Web GIS changed [2][7]. Then, Web GIS evolved from web browsers to a web GIS that served desktop and mobile clients in addition to web browsers. This distributed information system has a server and a client where the server is a web application server and the client is a mobile application, a desktop application, or a web application server. The server contains an URL, so it can be easily reached by users [2]. The client sends a request and receives a response from the server. Today, web GIS refers to a type of any GIS that uses the web [7].

Web GIS is closely related to Internet GIS and geospatial web. However, they are slightly different. The relation between GIS and web GIS is shown in Fig. 1 [7]. The Internet supports many services; web is only one of them. A GIS that uses a service other than a web service on the Internet is considered Internet GIS. Web GIS is more frequently used on the Internet, which means web GIS is the most pervasive of the Internet GIS [2][7].

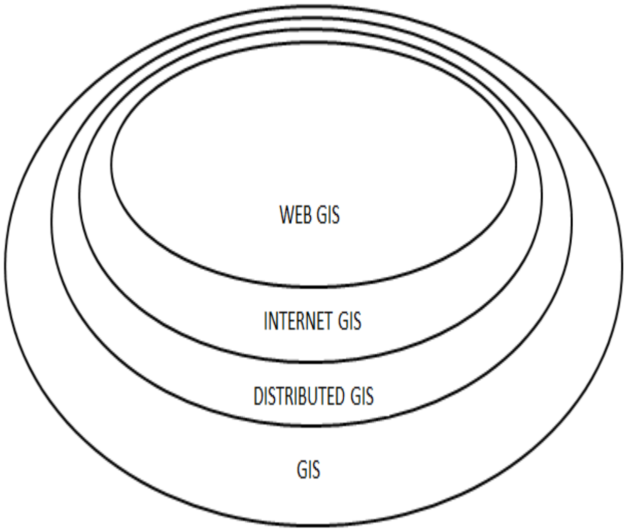


Figure 1. Types of GIS.

The geospatial web (geo web) is another term used instead of web GIS [2][7][13]. The geo web is slightly different from web GIS. The geo web merges geospatial information with non-geospatial information [2][13]. Another explanation is that the emerging geo web distributes global GIS, collaborates on information and knowledge with widespread worldwide sharing and interoperability [2][8][13]. This gives the idea that the geo web may be more developed in the future than web GIS to achieve the whole requirements.

The main characteristics of Web GIS are :global reach, a large number of users, better cross-platform capability [2][7][13] (because all web browsers use standards), low cost average by the number of users, easy to use by end-users, unified updates, and diverse application based on the use region. These characteristics create advantages, but also challenges. Web GIS stimulates public participation, but users with no GIS background must be considered. The increase in user numbers must be considered to provide good service based on the users' hardware, connection, and software limitations or capacity. These issues must be considered for appropriate communication or use.

Web GIS [7] can provide all GIS functions but also has unique strengths. These strengths include collecting geospatial information, mapping and querying functions, disseminating and analyzing geospatial information. An example of collecting geospatial information is an open street map formed by volunteers who collect data and share them in different parts of the world. This includes volunteered geographic information (VGI) [7]. We do not provide the details in this study as it is beyond the scope of the paper.

Web GIS is a new business model used by companies (the business location and the sponsor location of web sites on the map advertise them). It is a powerful tool for e-government. They have grown since 1993 and web GIS has grown more, but the fact that they engage the powerful representation and analysis capability of the GIS [7] makes things on online maps easy to understand. Another use is the infrastructure for e-science, which provides a low-cost, newly established readily accessibility infrastructure for computing capability. Web GIS [2][7] uses a component of daily life with the spread of mobile devices and the mobile web. Web GIS helps us in everyday life, by telling us where to go, what to do, where to eat, and how to get there based on LBS.

C. Mobile GIS

Mobile GIS refers to mobile use of the GIS [2][7]. A special aspect of mobile GIS is when data users are in someplace and are interested in information about a particular location. There are four types of locations in a GIS [2]: the user's location (u), storage location of data (d), the location where it is been processed (p), and the area subject to the GIS project(s). In a traditional GIS, the user sits at the location and projects to the world in which (u) not equal to (s). However, in mobile GIS u=s, which means the user is located at the projected area [7].

In the 1990s, mobile GIS were used for vehicle navigation and field surveying. GIS data were obtained from preloaded data in the system with a disconnected mode [9]. Then, with the advantages of wireless technologies, a GIS application is instantly connected to the Internet in real time to gather or view data [2]. These applications use web GIS services over the Web, which are advantages of the latest updates of the data and post the new data on the application to the server. That means that over the years mobile GIS has evolved to connect online to the resources and become a component of web GIS [7]. Mobile GIS overlaps with some parts of web GIS.

Since the use and popularity of mobile devices have increased, the increasing speed of data transfers by wireless networks has forced users to adopt the new form of web GIS, which is mobile GIS. Mobile GIS functions are quite compatible with some GIS functions. The following functions work well with mobile GIS [2][7]: information capture and updating, dissemination, storage, analysis, and presentation. Based on the types of applications, there are two mainsubtopics: consumer mobile GIS application and enterprise mobile GIS application.

The consumer mobile GIS application answers general daily life requirements, which mainly refers to location services based on LBS [2][10] and provides added value such as points of interest. Enterprise mobile applications are used to complete tasks such as field mapping queries and decision support, field inspection and inventory of assets, field surveying, incident reporting, collaboration, and tracking. Enterprise applications have more functions than consumer applications.

The advantages of using mobile GIS instead of a paper application are mobility, large volume of users, location awareness, versatile means of communication, and near-real-time information [7]. Supporting technology of these types of application include mobile phones, pocket PCs, portable PCs, and special devices such as operating system embedded or GPS receiver-embedded mobile devices. Mobile positioning [7] techniques used in mobile GIS include the navigation satellite-based approach, cellular-network-based approach, cellular-network-based approach, assisted GPS approach, Wi-Fi based approach, and IP address-based approach. In addition to mobile approaches some use different types of browsers such as mini browsers, full html, mobile browser plug-ins, and WAP.

D. Open Spatial Consortium

The OGC is an international industry consortium of 482 companies, government agencies, and universities that make sure publicly available interface standards are developed [2][8].“OGC® Standards support interoperable solutions that ‘geo-enable’ the Web, wireless, and location-based services and mainstream IT” [8]. The standards empower technology developers to make complex spatial information and services accessible and useful for all types of applications.

The standards are developed in unique consensus supported by the OGC's industry, government, and academic members [8]. Members want to make certain the geoprocessing technologies interoperate, or are "plug and play.” Products based on OGC standards receive the OGC certificate [8]. For interoperability and standardization issues, if people want to be in the market or create products that work well and are compatible with other products, they must follow OGC standards.

OGC standards are the documents established by consensus and accepted by members that have achieved the optimal degree of interoperability [8]. In 1998, the first standards were created; in 2000, the first web service standard was created; in 2004, the name was changed into to its current name; and today, there are 35 standards and 482 members [8].

The most important OGC standards are the sensor web standards [8], a type of sensor network. Many people think it is part of information technology, but the sensor network is in the geospatial domain. The main topic areas for which the OGC creates standards include cryoscopy, hydrology, meteorology/oceans, defense, intelligence, sensor web, aviation, and 3D modeling and visualization.

When the OGC was first started in 1994, the organization was interested only in creating standards for distributed GIS data and software interoperability [8]. Today, open GIS web coverage service, open GIS web processing service, open GIS service model implementation standard, open GIS georeference table joining standards, and geography mark-up language (GML) [2][8] standards are examples of standards created by consensus of the OGC.

E. Ubiquitous GIS

Ubiquitous GISs are based on a type of geographic information provided to users or systems at any time and any place through communication devices when the distributed system was formed [2][12][14]. Ubiquitous GIS (UBIGIS) is closer to the aim that people access data and processing them at anywhere and anytime. The information provided for the user is based on the user's context. The GIS has contextual awareness. It includes a set of practices and standards for spatial and geographic information which is accessible in public by users. Its importance is that UBIGIS also processes the applications.

According to some researchers, four criteria must be applied ("distributed," "disaggregated," "decoupled," and "interoperable") for UBIGIS [12] [14]. In computing area and distributed GIS, interoperability is the most difficult issue; it is negotiated by the OGC. UBIGIS applications must adhere to the following standards:support applications, numerous users, collaborative work, working online and off-line, multi-functional, security, and integration with other applications and various networks [11][12][14].

The future development of mobile GIS in many disciplines and application areas are toward ubiquitous GIS which are: augmented reality, public participation GIS, dynamic demography, and 4D GIS [7] [12]. These topics are more related to ubiquitous GIS. Augmented reality combines information from databases with information from the senses. This is implemented in medical settings where mobile GIS are used instead of the senses [7][14]. For example, visually impaired people navigate with mobile GIS. Public participation GIS (PPGIS) makes each mobile phone owner as being a sensor [7]. This allows people to add data to the network collectively in real time. In addition, real-time environmental monitoring networks harvest valuable data from the intelligence of public participation. Dynamic demography is for the planning of traffic and transportation choosing retail store place and simulating diffusion of disease based on these demographic data. A four-dimensional GIS performs by collecting mobile GIS data in real time for 4D [7]. For faster networks with 4G, the GIS must become more ubiquitous for users.

F. Sensor Network

A wireless sensor network is a geographically distributed network that monitors physical or environmental conditions. Sensor networks are formed by sensor nodes called "motes" [15]. Motes ease communication between each other [15]. They are consisting of radio transceivers which also contain memory, onboard power supplies and a variety of sensors. The motes work by collecting and analyzing the sensor readings independently but can also link up with neighboring motes with mesh topology to send information to each other or to the sink [15]. There are different applications in different areas and many more details; however, in this paper, we are more interested in the sensor web. This is a type of wireless sensor network based on automatically communicating sensors and reacting to phenomena. The sensor network is defined after the sensor web.

A sensor web is a system that is wireless, an intercommunicating spatially distributed sensor that is deployed to monitor and explore the environment [8][16]. It is capable of automated reasoning, and it responds to changing environment conditions and carries out automatic recovery after automatically sensing a diagnosis. OGC-defined protocols create abstract heterogeneous communication inside or between different networks by abstracting the basis and communication standards of the sensor web [8][16]. Sensor Web Enablement (SWE) standards have been created by the OGC [7][8]. It is also related to the communication protocols and types of the hardware. Defining accessing sensor network is in real-time or archived data using standard protocols and application programming interfaces [8]. This issue is starting to relate to the "Internet of Things" issue, in which real object world is directly related to the Internet. The sensor web is a network of sensors accessible to the sensors, and stored old data can be discovered using the OGC standard protocols and APIs [8].

G. Internet of Things

The Internet of Things was first mentioned as identifiable objects in the virtual world, on the internet places with tags [17]. The Internet of Things is a trending topic today, which means real world objects communicate with each other from the base on the Internet directly; the sensor web is also part of this relation. It is really a new area to improve nowadays and is expected to be fully constructed by 2020. Protocols and key applications were introduced to us a few years ago, and they are developing now. However, technical details are beyond the scope of this paper. Key applications used today include a sensor web used for intrusion detection or environmental monitoring. Various organizations are working in this area and have proposed objects with auto IDs to communicate with each other over the Internet. Radio frequency Identification (RFID) tags are a prerequisite of the Internet of Things [21].

The most important application of the Internet of Things is the smart grid, an electrical grid that uses information and communications technology to obtain and use information, such as information about suppliers and consumers, in automated processes to improve the efficiency, reliability, economics, and sustainability of electricity production and distribution [18]. Another important application is electric vehicle charging, for which the standards are being determined. Two critical standards that have been introduced are for plugs and charging stations for extended use with one charge later, so different areas can charge different vehicles [18].

Since this area is still being developed, we will see clearer details and more applications in the future. In the next 10 years, many applications will be distributed and affect our daily lives with this application. Application fields include waste management, emergency response, intelligent shopping, smart product management, smart meter, and home automation. Several applications are not distributed, such as the home automaton application sensor network, but they are distributed around the house and provide information about the house that is connected to other processes that may act automatically.

VI. Future GIS Trends and Research Areas

Future GISs research topics include VGI, PPGIS, geocollaboration, geotagging, geoparsing, geotargetting, online VR, Digital Earth, and cloud-based GIS. Future trends include faster and more mobile Internet and a smarter and more sociable web.

VGI is geospatial data about the environment generated by users [7] instead of producers. Examples include wikimaps, (Picasa, Panoramio, Flickr albums (geotagging)), OpenStreetMap, and Facebook (for location tagging). It provides public participation in GIS as people act as sensors, using their free time to provide data. Usually, producing data requires training, but since there are many data, we can choose the valuable one. In addition, people's adding data voluntarily will also help improve PPGIS [7][20].

PPGIS is people participation in GIS. It is a type of VGI, but requires more people to add data. PPGIS differs from VGI because participation is needed in decision making through the use of GIS [7][20]. PPGIS wants public participation in satellite image, digital maps, and sketches. However, PPGIS has technical and social barriers. Some people not want to attend public organizations at specific times, and some are not comfortable speaking at public forums. Some people may find it hard to use these technologies. Most current PPGIS applications are experiments or pilot studies [20]. The plan is to involve the public in PPGIS in a web-map-based approach with specific comments with environmental areas, blog and photo sharing web sites. A social networking approach may help improve PPGIS.

Geocollaboration also called collaborative GIS, concerns a group of people working together on the same task [7]. Two types of spatial collaboration talk about the same location at different places looking at the same part of the map, the second one temporarily can be synchronous or asynchronous [7]. Five characteristics must be completed: facilitating dialog, accounting for group behavior, drawing the group's attention, allowing private work, and allowing saved and shared sessions. The possibility of increasing geocollaboration is promising for the future.

Geotagging involves adding location information to photos, videos, or other digital data [7]. This can be done automatically and manually. Photos taken with devices with GPS receivers are automatically geotagged, and others are manually geotagged. Geotagging provides value to a GIS by spatially organizing photos and other types of data, enriching the geographic information, and providing valuable data for data mining.

Geoparsing is the process of converting geographic locations to textual words or phrases in a document [7]. It is a new research area and can be linked to the semantic web [7]. Geoparsing searches documents quickly, plots locations on the map easily, and organizes a mass of documents spatially.

Geotargetting determines the physical location of the user and provides related content based on that information. An example is advertising swimsuits in hot regions and near sea locations and advertising warm coats in colder regions [7]. Methods for obtaining physical locations include obtaining user's registration information or a user's IP address or from GPS receivers in mobile phones. Geotargetting helps provide precise advertising based on location, prevents suspicious payments, and by legal and license regulation online services or products are delivered only certain countries and regions.

Virtual reality (VR) interacts with the real world virtually to examine or do things [7]. This provides vividness and interactivity. Online VR can rotate various 3D objects (Earth), viewing 3D structures (such as buildings you can see inside before you go), and high-end specialized VR, which is used in game consoles to make players feel as if they are inside the game.

Digital Earth is partly created with 3D mapping, Google Earth, or NASA World Wind, but the real target is eight key elements for Digital Earth: multiple connected digital earths for different needs of audience, problem oriented, allowing the search of the time and space to find similar situations, asking questions about changes, enabling complex data and analysis services, supporting visualization of abstract concepts and data types, multi-disciplinary education lab, and open access for the multiple different platforms and technologies [7][22].

Cloud GIS is GIS software and services on a cloud infrastructure and accessing GIS capabilities using web services. Cloud GIS is a computer paradigm for on-demand network access shared facilities [7][19]. Cloud computing offers an alternative method for GIS software and services to users or customers. Instead of running a GIS on computer systems, with a cloud GIS, the software and services reside on cloud servers mainly (not every time) and are made available through web technologies [19]. With the help of cloud computing, powerful tools are not needed because work may do on the server. However, a disadvantage is that the services may not be available all the time. A cloud GIS may reduce the cost of buying hardware or software to deposit data.

For the future trends, faster and more mobile Internet will be based on the development of a prototype being tested in the USA: 100 Gbit Internet. If the work goes well, this improved area will also affect the GIS. More IP addresses may provide IDs for all objects in the Internet of Things with help of IP6. More mobile is based on faster mobile network technology with the impact and tests conducted with 4G networks. Using more collective intelligence, lightweight programming models, software running on multiple devices, open geospatial web service, making more powerful web clients, mobile serving the pervasive platform, more intelligence on web GIS, and serving GIS directly from clouds are future research areas.

VII. CONCLUSION

The aim of this paper was to summarize the important concepts and main points of the GIS in the pervasive area. The main concepts of the GISs which are very important last few years, provided information about application areas, and compared some areas with each other. GIS use is based on the user's conditions and is application specific. In addition, GIS is a growing area. Many books and materials explain GIS in technical detail and provide examples. Mainly GIS challenge continues because of the technological evolution. This paper outlined the GIS evolution up to participation in the pervasive area, summarized the main points, and examined the beginning and current states of GIS. Mainly to sum up the work the internet effect on the GISs are positive and make the technology involve more our daily lives. Also the evolution continues, so we see many new things couple of years and these technologies growing and more important next years. As the challenge increases it is positive effect on the people. To conclude distributed and wireless systems make everything become easier than ever before, also examined new research concepts and future trends and suggested GIS trends that should be investigated in the future.

REFERENCES

[1] Longley, P. (Ed.). (2005). Geographic information systems and science. John Wiley and Sons.

[2] Peng, Z. R., and Tsou, M. H. (2003). Internet GIS: distributed geographic information services for the internet and wireless networks. John Wiley and Sons.

[3] Worboys, M. F., and Duckham, M. (2004). GIS: a computing perspective, CRC press.

[4] Geographic Information Systems (GIS) retrieved: January 2014.
http://www.esri.com/~/media/c371f47805c345fa84d32ac8a675046e.pdf

[5] History of GIS , retrieved January 2014. http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/u23.html.

[6] Tait, M. G. (2005). Implementing geoportals: applications of distributed GIS.Computers, Environment and Urban Systems, 29(1), 33-47.

[7] Fu, P. and Sun, J.(2010). Web GIS: principles and applications.Esri Press.

[8] Open Geospatial Consortium, retrieved: January 2014.
http:// www.opengeospatial.org

[9] Lemmens, M.(2011). Geo-information: technologies, applications and the environment (Vol. 5). Springer.

[10] Jiang, B., and Yao, X.(2006). Location-based services and GIS in perspective. Computers, Environment and Urban Systems, 30(6), 712-725.

[11] Weiser M.(1994, March). Ubiquitous computing. In ACM Conference on Computer Science (Vol. 418).

[12]Mnzis, A. H. ,and Msc, C.(2000). The Road to Ubiquitous The Geographic Information System s Roam Anywhere-Remain Connected.

[13] Scharl, A. and Tochtermann, K.(2007). The Geospatial Web: How Geobrowsers, Social Software and the Web2.0 Are Shaping the Network Society. Springer,

[14] Posland, S.(2011). Ubiquitous computing: smart devices, environments and interactions. John Wiley and Sons.

[15] Karl, H., and Willig, A.(2007). Protocols and architectures for wireless sensor networks, John Wiley and Sons.

[16] Di, L.(2007, February). Geospatial sensor web and self-adaptive Earth predictive systems (SEPS).In Proceedings of the Earth Science Technology Office (ESTO)/Advanced Information System Technology (AIST) Sensor Web Principal Investigator (PI) Meeting, San Diego, USA (pp.14). http://esto.nasa.gov/sensorwebmeeting/papers/di.pdf.

[17] Tan, L., and Wang, N.(2010, August). Future internet: The internet of things. In Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on (Vol. 5, pp. V5-376). IEEE.

[18]Hersent, O., Boswarthick, D., and Elloumi, O.(2011). The Internet of Things: Key Applications and Protocols. John Wiley and Sons.

[19] Bhat, M.A., Shah, R. M., and Ahmad, B.(2011).Cloud Computing: A solution to Geographical Information Systems (GIS). International Journal on Computer Science and Engineering, 3(2).

[20] Roche, S., Mericskay, B., Batita, W., Bach, M., and Rondeau, M. WikiGIS Basic Concepts: Web 2.0 for Geospatial Collaboration.Future Internet 2012, 4, 265-284.

[21] That 'Internet of Things', retrieved January 2014. http://www.rfidjournal.com/articles/view?4986.

[22] NASA World Wind retrieved January 2014. http://worldwind.arc.nasa.gov/index.html

Mapping Grasslands Formations and Cultivated Pastures in the Brazilian Cerrado Using Data Mining

Wanderson Costa, Leila Fonseca, Thales Körting

Image Processing Division
National Institute For Spatial Research
São José dos Campos, Brazil
{wscosta, leila, thales}@dpi.inpe.br

***Abstract*—Cerrado is the second largest biome in Brazil. Among the land changes in the Cerrado, over 500,000 km^2 of the biome have been transformed into cultivated pastures in recent years. Distinguishing the native formations and identifying types of land use and cover in the Cerrado are important tasks for monitoring and protection policy of the biome. Within this context, this work aims at developing a methodology based on remote sensing techniques, to map pasture and native grassland areas in the Brazilian Cerrado. Data related to land use and cover, relief, spectral information from Landsat images and vegetation indices were used to perform the image classification. Decision trees and Support Vector Machines algorithms were used, and the results showed that the analysis and integration of different data sources can aid in the classification process. In order to discriminate areas of cultivated pastures and grassland formations, we obtained accuracies up to 82% in the study area, being able to identify attributes and data required to recognize these areas in the Brazilian Cerrado by remote sensing images.**

Keywords-Image Processing; Data Mining; Cerrado.

I. INTRODUCTION

Serious environmental risks can arise and pose a threat of global character when natural resources are not used properly [1]. Within this perspective, it can be mentioned the problems caused by changes in land use and land cover in the second largest biome in Brazil, the Cerrado [2]. More than half of Brazilian Cerrado's area has been transformed, mainly to make room for cattle and cash crops, losing more than 1,000,000 km^2 of its original vegetation [3].

Croplands cover more than 100,000 km^2 and pastures surpass 500,000 km^2, whereas protected areas comprise only about 33,000 km^2. The destruction of the Cerrado vegetation is three times larger than the amount of the deforested area in the Amazon region. Deforestation ranges from 22,000 to 30,000 km^2 per year, even higher than those observed in Amazon Forest. This scenario represents a high environmental cost and implies loss of biodiversity, soil erosion, vegetation degradation, water pollution, instability of the carbon cycle and probable regional climatic changes and variations in fire events, which are typical of the biome [4].

There is a large number of definitions for Cerrado and, as a result, several proposals for classification of native formations can be found in the literature. Among the physiognomic types presented in the biome, there are the grassland formations, which include the physiognomies of *Clean Field*, *Dirty Field* and *Rocky Field* [5]. We argue that with the identification and monitoring of these vegetation types using satellite imagery, policies can promote its physical, chemical and biological integrity, and estimate the productivity of the degraded regions [6].

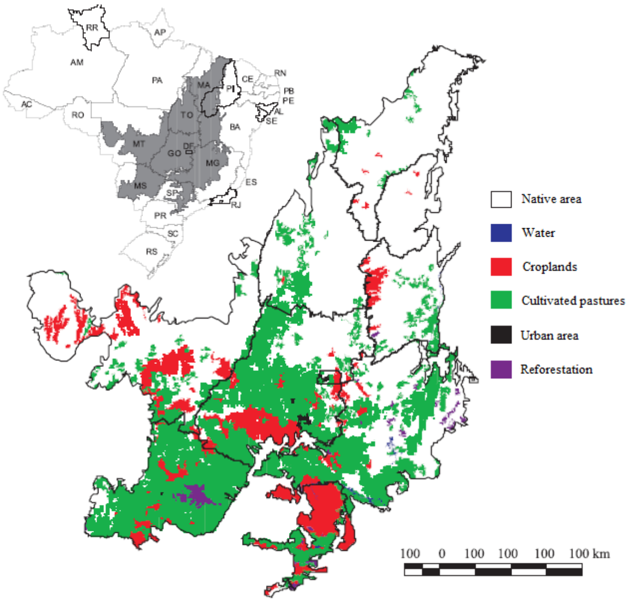


Figure 1. Spatial distribution of cultivated pastures in the Brazilian Cerrado biome. Source: [3].

In order to promote the recovery of degraded areas and the policies that protect the biome, it is essential to create maps to analyze the land use and land cover of the savanna. However, cultivated pastures, similarly to native grasslands, may vary from formations with predominance of grasses to areas that present dominance of pioneer trees and shrub species [7]. Therefore, the mapping of pasture areas and native formations in the Cerrado is a difficult task if only the spectral information obtained by satellite images is used [8].

To overcome this problem, this work proposes a methodology for mapping areas of Cultivated Pasture and Native Vegetation (Clean Field, Dirty Field and Rocky Grasslands), using data mining from integration of satellite imagery data in multiple resolutions.

The remainder of the paper is organized as follows. In Section 2, we present a brief description of the Brazilian

Cerrado. Section 3 describes the cultivated pastures in the biome. The methodological procedures and data used are depicted in Section 4. In Section 5, we discuss the results obtained in this work. Finally, we describe the conclusion and future work in Section 6.

II. The Brazilian Cerrado

Cerrado is the Portuguese word for Brazil's plateau of savannas, grasslands, woodlands, and gallery and dry forests [9]. One of the most biodiverse regions on the planet, the Brazilian Cerrado has an area of approximately 2 million km², comprising about 24% of the Brazilian territory [10]. Cerrado is the richest tropical savanna in the world [4] and the biome occupies the central region of Brazil, extending from the northeast coast of the Maranhão state (MA) to the north of Paraná (PR) state, as shown in Figure 1. Overall, Cerrado can be understood as a grass field coexisting with scattered trees and shrubs [11] and it is the second largest of Brazil's major biomes, after Amazonia [4].

The biome is characterized as one that presents three types of formations: forests, savannas and grasslands. Generally, the grassland formation refers to regions with predominance of herbaceous species and some shrubs, without the occurrence of trees in the landscape [5].

The vegetation included in the grassland formations comprises areas of Clean Field, Dirty Field and Rocky Field (Figure 2). Regions of Dirty Field and Rocky Field present a physiognomic type predominantly herbaceous and shrubby. However, Rocky Field areas include micro-relief landscapes with typical species, which usually occupy regions of rocky outcrops at elevations above 900 m. On the other hand, the phytophysiognomy of Clean Field has a predominance of grasses interspersed with underdeveloped woody plants, without the presence of trees [5].

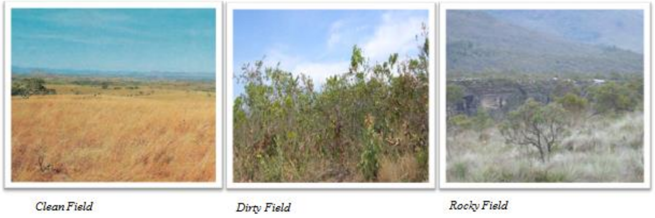


Figure 2. Grassland formations in the Brazilian Cerrado. Sources: [12], [13].

More than half of Cerrado's original vegetation has been transformed into cultivated pasture areas, agriculture and other uses, as shown in Table I. In addition, studies indicate that changes in land use in the Cerrado occur with greater intensity than in the Amazon region [14] [15].

TABLE I. Principal land use in the cerrado. Source: [4]

Land use	*Area (ha)*	*Percent core area*
Native areas	70,581,182	44.53
Cultivated pastures	65,874,145	41.56
Agriculture	17,984,719	11.35
Urban areas/bare soil	3,006,830	1.90
Planted forests	116,760	0.07
Others	930,304	0.59
Total	158,493,921	

Hence, these changes in land use impose substantial threats to ecosystems and species of the biome. Only 2.2% of its area is legally protected and various species of animals and plants are endangered [4]. Furthermore, approximately 20% of endangered endemic species no longer exist in preserved areas. Besides the vegetation degradation and soil erosion, the introduction of non-native species and the use of fire to create pastures can have an adverse impact on ecosystems and cause significant loss of biodiversity in the Cerrado [4].

III. Cultivated Pastures

Cultivated pastures represent about 500,000 km² of the biome. It is important to analyze the degradation of cultivated pastures, since nearly 50% of the planted areas in grasslands are severely degraded (Figure 3), causing increased erosion, loss of soil fertility and the predominance of invasive species [16]. Therefore, not only can the recovery of these areas increase the producers' income, but also it can reduce the environmental impact by decreasing erosion, emission of carbon dioxide and opening new areas for cattle [17].

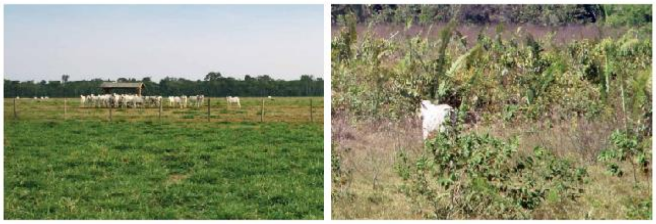


Figure 3. Examples of managed (left) and degraded pastures (right). Source: [7].

Some types of data can be extracted and analyzed to support the detection of these regions, such as biophysical (soil type, percentage of green cover and biomass), radiometric (Enhanced Vegetation Index – EVI and Normalized Difference Vegetation Index – NDVI) and climatic (precipitation) data. Moreover, wet and dry seasons are well defined in the region and, therefore, this fact may assist to identify the radiometric and biophysical characteristics of planted pasture areas [16].

The classification of cultivated pastures is difficult because the degradation of these regions can, for example, influence the percentage of vegetation cover and the response of vegetation indices. Misclassification may occur when pastures are managed improperly, since those areas might be detected as invasive species or even the revival of species of native shrubs and trees in these regions [16].

Therefore, in order to improve the discrimination of such targets it is necessary to use temporal and field data and also to better understand the biophysical properties of these areas. [8]. Thus, the analysis using images with different spatial and temporal resolutions can allow the more accurate identification of spatial and temporal patterns of the targets in the Cerrado.

IV. DATA AND METHODS

This study encompasses a Cerrado area that comprises a region of Serra da Canastra National Park and neighboring regions (Figure 4). The region is located in the south-central state of Minas Gerais, southeastern of Brazil. The chosen scene contains the targets of interest, both native grassland areas and cultivated pasture.

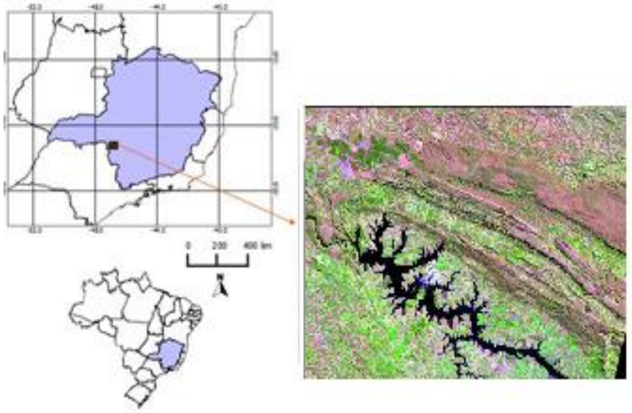


Figure 4. Location of the study area.

Information on cultivated pasture areas, for the year 2006, were provided by the Brazilian Ministry of the Environment. The regions of native areas, for the year 2009, were obtained from the Forest Inventory of Minas Gerais, created by the University of Lavras (UFLA), which ranks the grassland formations into two classes: Field (Clean and Dirty Fields) and Rocky Field. As a means to evaluate the classification results, these reference data were used as ground truth.

Therefore, the types of classes used in this work includes the following standards: Cultivated Pasture, Field and Rocky Field. The experiments were performed in the system GeoDMA [18]. This system performs image segmentation, classification, temporal analysis and data mining, and it is available on the internet.

Data related to relief, spectral information from Landsat images and vegetation indices EVI2 (Enhanced Vegetation Index 2) were used to perform the classification. For the classification of the areas, this work used the algorithms of decision tree C4.5 [19] and Support Vector Machines (SVM) [20], [21], using the system GeoDMA and the tools Weka [22] and LibSVM [23].

V. RESULTS

In order to implement an efficient methodology for distinguishing cultivated pastures and native vegetation, several experiments were performed. In the first experiment, it was used the Decision Tree (DT) classifier and only information about topography, obtained from altitude and declivity, called TOPODATA [24], with a spatial resolution of 30 meters, to map four classes: Cultivated Pastures, Field, Rocky Field and Others. The class Others covered all other native formations in the study that are neither Field nor Rocky Field. The results with 8,000 samples showed that regions below 871.47m were classified as Cultivated Pasture and areas above this value were classified as native vegetation. However, the accuracy of the results was only 60.0%.

In the second experiment, two Landsat-5 TM images (spatial resolution of 30m) were used for the orbit/point 220/74, acquired on 06/01/2006 and 07/27/2009. These images were georeferenced from NASA database (Global Land Cover Facility – GLCF). Various spectral attributes, such as mean, amplitude, homogeneity and entropy were extracted for bands 1-5 and 7. These were combined with the relief data previously used to classify the four classes. The classification accuracy based on decision tree slightly improved to 63.50%.

Because of the confusion between Field and Rocky Field classes, these were combined into a single class. For 12,000 samples, the results using the DT algorithm increased the accuracy to 71.65%. Finally, the SVM was also tested on the same set of samples, and the classification accuracy was 73.65%.

In the last experiment, it was added to the previously used set of attributes the time series of images EVI2 MODIS sensor (250m spatial resolution), with an interval of 16 days for the years 2006 and 2009. In addition to measuring accuracy, Kappa Coefficient [25] showed significant values in this experiment. Using the typology of four classes, the decision tree resulted in 60.15% of accuracy and Kappa equal to 0.468, while an accuracy of 67.20% and Kappa of 0.563 for SVM. With the combination of classes Field and Rocky Field, the accuracy increased to 73.27% (DT) and 82.12% (SVM), as well as Kappa Coefficient, that corresponded to 0.599 (DT) and 0.7319 (SVM). Table II summarizes the experiments performed so far.

TABLE II. EXPERIMENTS

Data used	*N. of classes*	*Accuracy (%)*	*Algorithm*
TOPODATA	4	60.00	DT
TOPODATA + Landsat	4	63.50	DT
TOPODATA + Landsat	4	71.65	DT
TOPODATA + Landsat	3	73.65	SVM
TOPODATA + Landsat + EVI2	4	60.15	DT
TOPODATA + Landsat + EVI2	4	67.20	SVM
TOPODATA + Landsat + EVI2	3	73.27	DT
TOPODATA + Landsat + EVI2	3	82.12	SVM

As shown in Table II, by the end of this stage of work, the experiment that combined different spatial resolutions (TOPODATA, Landsat and EVI2) and the SVM algorithm obtained the best classification, with an accuracy of 82.12% and Kappa of 0.7319.

VI. CONCLUSION AND FUTURE WORK

In order to develop a methodology to combine image processing and analysis to identify areas of pasture and grassland formations of the Brazilian Cerrado, some experiments were performed. The best classification result was the one that combined different spatial resolutions and the SVM algorithm, with an accuracy of 82.12% and Kappa of 0.7319, distinguishing areas of cultivated pasture and native grassland.

However, further experiments will be carried out, by integrating images from different sources to aid in the characterization of targets. Images of high spatial resolution, as well as techniques of Linear Spectral Mixture Model and principal components will be tested to evaluate the most proper features for each of the targets.

References

[1] J. Beddington, “Food, energy, water and the climate: A perfect history of global events?,” *Lecture to Sustainable development UK 09,* 2009, pp. 1-9.

[2] J. A. Ratter, J. F. Ribeiro, and S. Bridgewater, “The Brazilian Cerrado and threats to its biodiversity,” *Annals of Botany,* vol. 80, 1997, pp. 223-230.

[3] R. B. Machado et al., “Estimated loss of the Brazilian Cerrado area” Brasília, DF, 2004, pp. 1-26.

[4] C. Klink and R. Machado, “Conservation of the Brazilian Cerrado,” *Conservation Biology,* vol. 19, no. 3, Jun. 2005, pp. 707-713.

[5] J. F. Ribeiro and B. M. T. Walter, “The main phytophysiognomies in Cerrado,” in *Cerrado: Ecologia e Flora*, vol. 1, Brasília, EMBRAPA, 2008, pp. 152-212.

[6] R. R. Rodrigues and S. Gandolfi, “Recovery of Native Forests: General Principles and Allowances for a Methodology Definition,” *Revista Brasileira de Horticultura Ornamental,* vol. 2, no. 1, 2001, pp. 4-15.

[7] Embrapa and Inpe, “Survey information of use and land cover in the Amazon,” Brasília, 2011, pp. 1-20.

[8] L. G. Ferreira, E. E. Sano, L. E. Fernandez and F. M. Araújo, “Biophysical characteristics and fire occurrence of cultivated pastures in the Brazilian savanna observed by moderate resolution satellite data,” *International Journal of Remote Sensing*, vol. 34, no. 1, 2013, pp. 154-167.

[9] G. Eiten, "Delimitation of the concept of Cerrado," vol. 21, Rio de Janeiro: Arquivos do Jardim Botânico, 1977, pp. 125-134.

[10] M. Brossard and A. O. Barcellos, “Cerrado conversion into cultivated pastures and operation of latosols,” *Cadernos de Ciência & Tecnologia,* vol. 22, no. 1, 2005, pp. 153-168.

[11] B. M. T. Walter, “Phytophysiognomies of Cerrado biome: terminological synthesis and floristic relationships,” Brasília, 2006, pp. 1-373.

[12] IBGE, "Technical Manual of Brazilian Vegetation," 2 ed., vol. 1, Rio de Janeiro: Manuais Técnicos em Geociências, 2012.

[13] S. M. d. C. Coura, "Vegetation Mapping of Minas Gerais State Using MODIS Data," São José dos Campos, SP, 2006, pp. 1-150.

[14] E. E. Sano, A. O. Barcellos, and H. S. Bezerra, “Assessing the spatial distribution of cultivated pastures in the Brazilian savanna,” *Pasturas Tropicales,* vol. 22, 2001, no. 3, 2001, pp. 2-15.

[15] D. L. Skole, C. W. H. W. A. Salas, and C. A. Nobre, “Physical and human dimensions of deforestation in Amazonia,” *Biosciences,* vol. 44, no. 5, 2012, pp. 314-322.

[16] L. G. Ferreira, et al., “Biophysical Properties of Cultivated Pastures in the Brazilian Savanna Biome: An Analysis in the Spatial-Temporal Domains Based on Ground and Satellite Data,” *Remote Sensing,* vol. 5, no. 1, 2013, pp. 307-326.

[17] J. M. Chaves, L. Moreira, E. E. Sano, H. S. Bezerra, and L. Feitoza, “Using the segmentation technique to identify the main types of cultivated pastures in the Cerrado,” em *Simpósio Brasileiro de Sensoriamento Remoto, 10 (SBSR)*, Foz do Iguaçu, 2001, pp. 31-33.

[18] T. Korting, L. M. G. Fonseca, and G. Câmara, “GeoDMA - Geographic Data Mining Analyst: a framework for GIScience,” *Computers & Geosciences,* 2013, pp. 133-145.

[19] I. H. Witten and E. Frank, Data mining: practical machine learning tools and techniques with Java implementations, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000.

[20] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Maching Learning,* vol. 20, no. 3, Set. 1995, pp. 273-297.

[21] F. Bovolo, G. Camps-Valls, and L. Bruzzone, “A support vector domain method for change detection in multitemporal images,” *Pattern Recognition Letters,* vol. 31, no. 10, 2010, pp. 1148-1154.

[22] M. Hall et al., “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations,* vol. 1, no. 1, 2009.

[23] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transaction on Intelligent Systems and Technology,* vol. 2, no. 3, 2011, pp. 1-27.

[24] M. d. M. Valeriano, “Digital model of morphometric variables with SRTM data nationwide: the project TOPODATA,” *Anais do XII Simpósio Brasileiro de Sensoriamento Remoto,* 2005, pp. 3595-3602.

[25] J. A. Cohen, “Coefficiente of agreement for nominal scales,” Educational and Psychological Measurement, no. 20, pp. 37-46, 1960.

GIS- Interpolated Geotechnical Zonation Maps in Surfers Paradise, Australia

Haider Al-Ani, Erwin Oh, Gary Chai

Griffith School of Engineering
Griffith University
Gold Coast, Australia
E-mail: h.al-ani@griffth.edu.au,
y.oh@griffith.edu.au, g.chai@griffith.edu.au

Bahar Nader Al-Uzairy

Griffith School of Environment
Griffith University
Brisbane, Australia
E-mail: bahar-nader.al-uzairy@griffithuni.edu.au

***Abstract*—Due to the escalating cost of site investigation in Australia, geotechnical data contributes substantially to the cost of engineering projects. The GIS (Geographic Information System) has been used as a vital tool in civil engineering in recent years for a variety of applications. The subsurface conditions of Surfers Paradise (as a case study) have been examined in terms of soil stiffness by using GIS. The Spatial Analyst extension in ArcMap10 has been employed to develop zonation maps for different depths in the study area. Each depth level has been interpolated as a surface to create zonation maps for the Standard Penetration Test SPT-*N* value of the soil for each depth. The Inverse Distance Weighting (IDW) method in the Spatial Analyst shows better representation for these zonation maps with certain parameters among 8 interpolation techniques.**

Keywords-soil stiffness; interpolation; Spatial Analysis, Inverse Distance Weighting (IDW); zonation maps

I. INTRODUCTION

Due to the escalating cost of site investigations in Surfers Paradise, geotechnical data are expensive and contribute substantially to the cost of the geotechnical work. Due to budgetary constraints, small projects often overlook site characterisation [1]. The GIS (Geographic Information System) has been used as a vital tool in civil engineering fields in recent years for a variety of applications. Statistical models have been developed in GIS and are widely used to evaluate landslide hazards [2][3][4]. In addition, Atkinson and Massari [5] developed a linear modelling of the susceptibility of land sliding in the central Apennines in Italy. Furthermore, GIS has been utilised to develop zonation map production and to estimate if further precaution is required for a safer area in Turkey [4]. In their study, three Standard Penetration Test (SPT) N value zonation maps have been interpolated by using GIS Spatial Analyst IDW method based on the field data [4].

In fact, geotechnical characterisation of an area was an arduous task before GIS because of complexity of soil logs and their data representation [6]. Thus, the need for the GIS which transforms all paper work (hard copies) into digital forms to make data quickly accessed and easily analysed, is inevitable. Zonation maps have been used in various disciplines providing information on slope stability, seismic microzonation, groundwater quality, watershed, vegetation, and landslide hazard assessment. In the past, zonation maps had been produced by using traditional paper maps or contour maps to show the distribution of zones within a map. Nowadays, in geotechnical engineering, for example, the zonation maps have been produced to determine the suitability of foundations in the residential area by using GIS [4]. Four case studies have been researched by Hellawell et al. [7] to verify the benefits of using GIS in geotechnical engineering in the United Kingdom. These case studies are related to small-scale geotechnical projects. The outcome of this study indicated that GIS's output enhanced the analytical and technical range of these projects in comparison with traditional techniques and the high quality maps produced are comprehensible and popular with engineers.

The Standard Penetration Test (SPT) is an in-situ test widely used because it is simple, fast conducted and low in cost [8] which had been developed around 1972 [9]. This test is used to determine soil constitutive behaviour, such as stiffness, to characterise the soil type (sand, clay, or gravel), and to provide some correlation with other properties of soil. Around 85 to 90 percent of conventional foundations are designed using SPT in North and South America [9].

In this paper, zonation maps have been developed for the soil stiffness in Surfers Paradise, Australia (see Fig.1). The purpose of this is to characterise the soil in terms of soil stiffness and to predict new SPT-N values from the existing data. This is to reduce the cost of soil investigation or at the very least to have an initial concept of the soil strength in the study area. This paper consists of, after the introduction, the methodology which has been used to achieve the results followed by presenting the results and discussion, and then, the conclusions have been drawn.

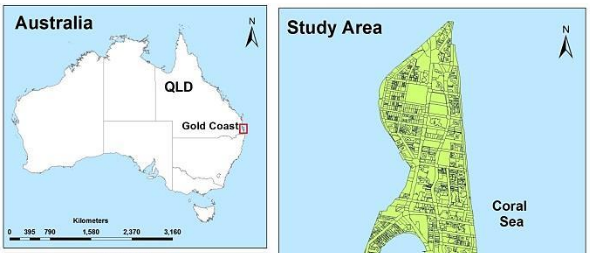


Figure 1. The study area location

II. METHODOLOGY

Data have been collected from 35 locations within the study area which extends about 1.3 by 4.0 km. Surfers Paradise is a business and tourism hub in Gold Coast city, Australia. The collected data includes 1,754 soil stiffness values which have been used as an input to the GIS. This data have been geo-referenced by obtaining the Easting and Southing coordinates from Google Earth and validated by using GPS for selected locations. In ArcMap10 and under the Spatial and Geostatistical Analyst extensions, 8 interpolation techniques have been examined in terms of their suitability to represent the SPT-N data. This has been done at the same depth which at Reduced Level (R.L. -11.6 to R.L. -13.3 m). The utilised interpolation techniques were (1) Geostatistical Analyst tools which include: Inverse Distance Weighting (IDW), Diffusion, Global Polynomial, and Kernel methods (2) Spatial Analyst tools which comprise: Ordinary Kriging, Universal Kriging, Spline, and IDW. The examination of these interpolation techniques has been done by attempting all the parameters of each method as an input to develop the SPT-N zonation maps (See Table I).

III. RESULTS AND DISCUSSION

26 zonation maps have been established from the existing Standard Penetration Test SPT-N value data by using the Spatial Analyst Inverse Distance Weighting IDW interpolation technique. Only four Standard Penetration Test SPT-N value zonation maps have been presented in this paper at depths between R.L. -1.6 m to R.L. -25 m at different depth intervals. In addition, a comparison has been done among 8 interpolation techniques to examine which technique provides better representation for the SPT-N value in the study area “see Fig. 2”. The Inverse Distance Weighting IDW interpolation technique provides better and reasonable representation for the SPT-N data among the aforementioned 8 interpolation techniques with certain parameters. These parameters are illustrated in Table II. Other interpolation techniques did not provide a correct representation. This has been discovered through matching the SPT-N values in the resulted zonation maps with the original input of the data where the resulted values was far away from the original input data. It shows that only the IDW technique provided a zonation maps with a resulted SPT-N values were very close from the original input data.

It can be seen from Table II that the power of the formula being used in the mathematical computations of the IDW technique is power 2. From the literature, Lloyd [10] utilised the power 2 of IDW to interpolate the precipitation values in the UK. In addition, Ping et al. [11] have used IDW power 2 in the exploring of spatial dependence of cotton yield in Texas.

TABLE I. THE EIGHT INTERPOLATION TECHNIQUES USED AND ITS RELEVANT PARAMETERS.

GIS Tools	Interpolation Technique	Parameters
Geostatistical Analyst	IDW	Output cell size, power, search neighbourhood, major semi axis, minor semi axis, max neighbour, min neighbour, angle
Geostatistical Analyst	Diffusion	Output cell size, number of iterations, weight field, band width
Geostatistical Analyst	Global Polynomial	Output cell size, order of polynomial, weight field.
Geostatistical Analyst	Kernel	Output cell size, Kernel function, order of polynomial, output surface type
Spatial Analyst	Ordinary Kriging	Output surface raster, semivariogram model (spherical, circular, exponential, Gaussian, linear), output cell size, search radius, number of points, maximum distance
Spatial Analyst	Universal Kriging	Output surface raster, semivariogram model (linear with linear drift, linear with quadratic drift), output cell size, search radius, number of points, maximum distance
Spatial Analyst	Spline	Output cell size, Spline type (regularized, tension), weight, number of points
Spatial Analyst	IDW	Output cell size, power, search radius (Fixed, Variable), number of points, Max distance

It has been also used for predicting and mapping of potassium soil by [12]. Other researchers have used the IDW method without mentioning to the power of this technique, such as [4]. Therefore, the value of power (2) which has been adopted in this research is considered according to [13], as a frequently used value.

Fig. 3 represents a GIS based zonation map for the SPT-N value at a depth of R.L. -1.6 to R.L. -3.2 m in Surfers Paradise. Most of the soil types at this depth have an N value of between 31 - 49 blows. This means that these soil types are dense sand, based on the soil classification given by Look [14] and based on the description provided in the soil investigation report of this location.

TABLE II. ADOPTED INVERSE DISTANCE WEIGHTING (IDW) PARAMETERS.

Method	Output cell size	Power	Search Radius	Distance
Spatial Analyst IDW	2.719E-05	2	Fixed	0.25

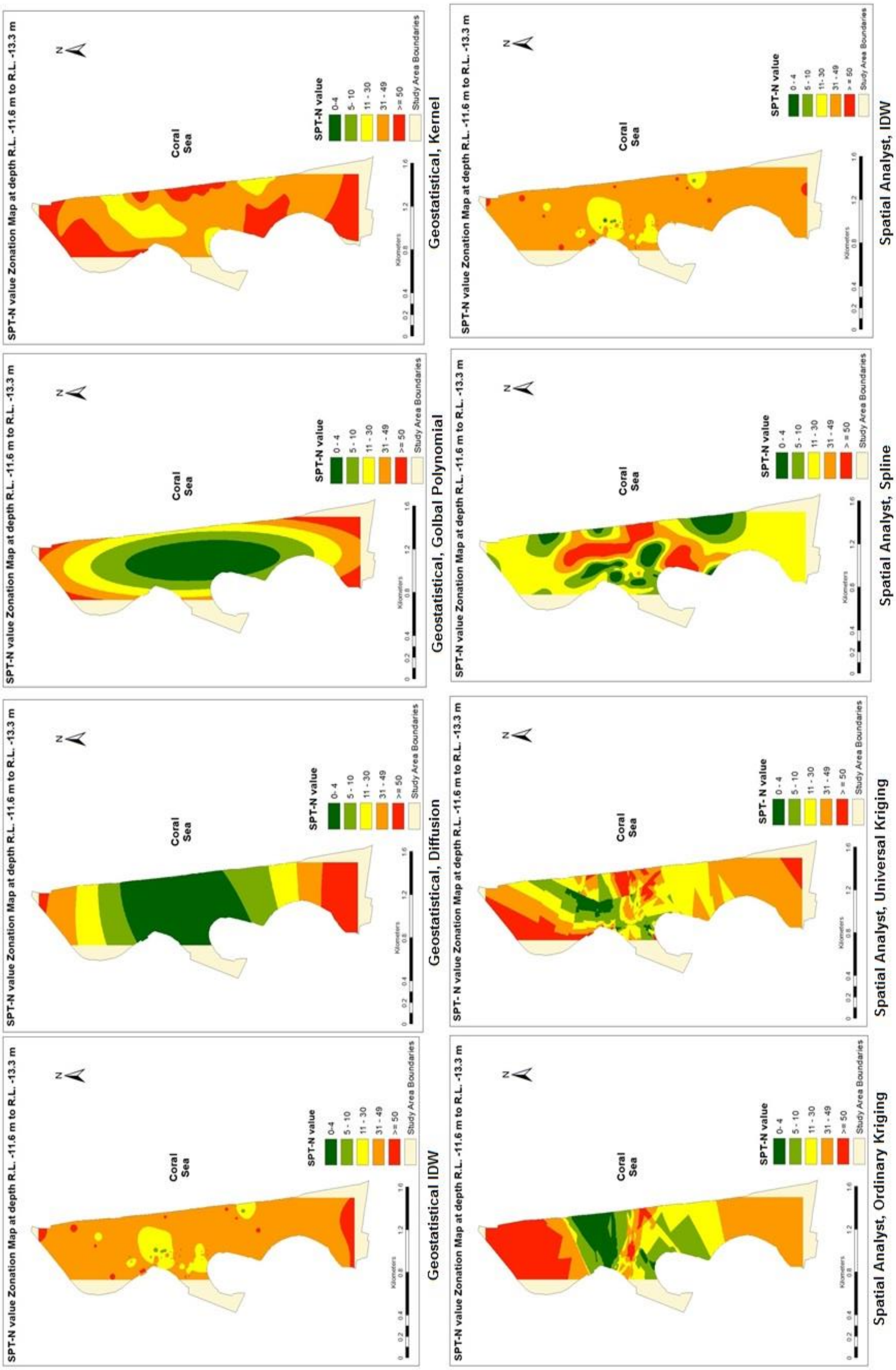

SPT-N value Zonation Map at depth R.L. -11.6 m to R.L. -13.3 m
SPT- N value Zonation Map at depth R.L. -11.6 m to R.L. -13.3 m
Coral Sea
N
SPT-N value
SPT- N value
0-4
5-10
11-30
31-49
>=50
Study Area Boundaries
Kilometers
0
0.2
0.4
0.8
1.2
1.6
Geostatistical IDW
Geostatistical, Diffusion
Geostatistical, Golbal Polynomial
Geostatistical, Kernel
Spatial Analyst, Ordinary Kriging
Spatial Analyst, Universal Kriging
Spatial Analyst, Spline
Spatial Analyst, IDW

Figure 2. Comparison among eight interpolation techniques in GIS

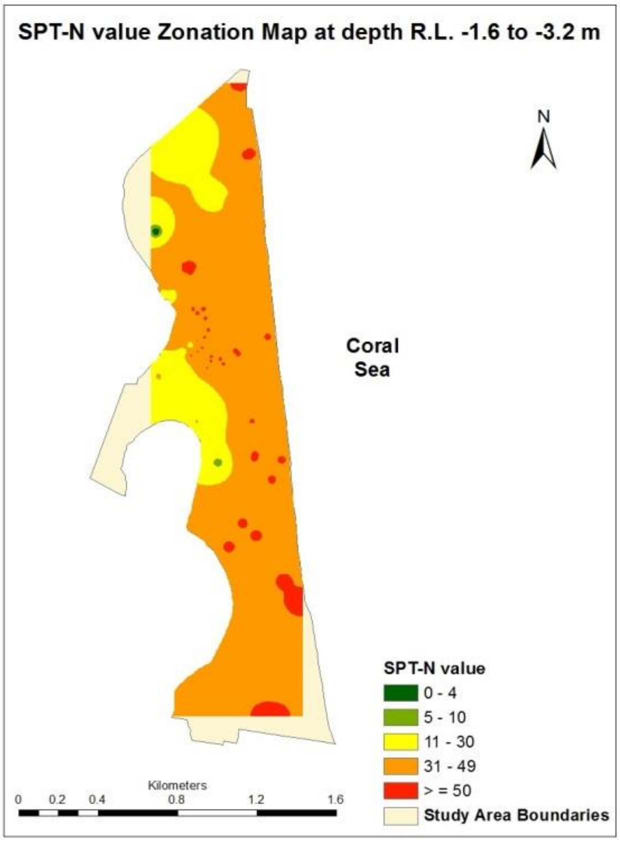


Figure 3. SPT-N zonation map at depth R.L. -1.6 to R.L. -3.2 m

However, there are some locations have an N value of between 0 to 4 and between 11 to 30 blows in the eastern parts of the study area. This reveals the occurrence of peat and medium dense sand in those locations respectively. In addition, there are well defined scattered red points throughout the map showing the occurrence of very dense sand with an SPT-N value of more than 50 blows.

Fig. 4 shows a GIS based zonation maps for the SPT-N value in Surfers Paradise at the depth of R.L. -3.2 to R.L. -5.0 m. It can be seen from this zonation map that the majority of the soil types at this depth has an N value of between 31 and 49 blows. This represents dense sand with an embedded peat and loose sand pocket in 5 different locations to the east of the study area. Also, many places have an N value of more than 50 blows which represent very dense sand.

It can be seen from Fig. 5 that the dense sand is dominant at this depth level with an N value of between 31 and 49 blows. There are some areas that have not been interpolated due to the lack of data in these locations. Further, other areas which have an N value of between 11 to 30 blows represent stiff clay to medium dense sand with few very dense sand locations in the south and north east of Surfers Paradise.

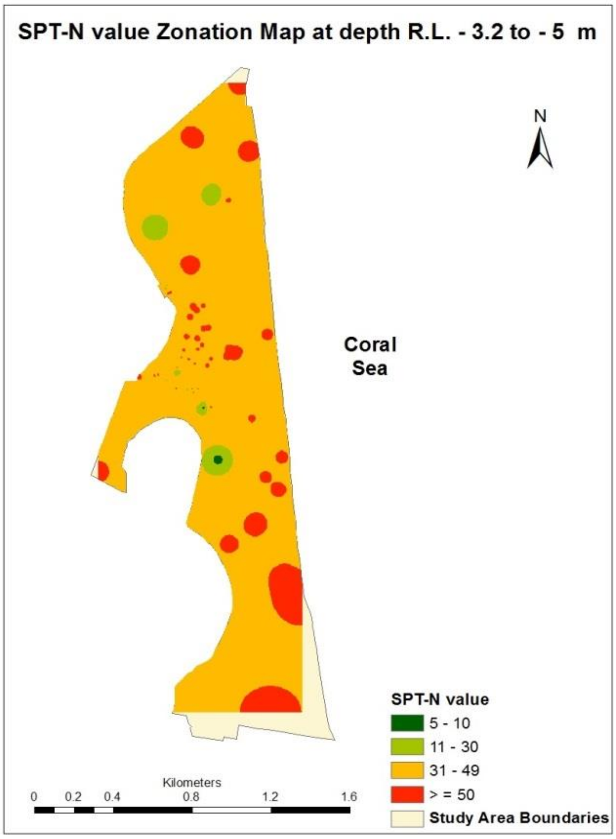


Figure 4. SPT-N zonation map at depth R.L. -3.2 to R.L. -5.0 m

Fig. 6 shows a GIS based zonation map at a depth of between R.L. -23.3 to R.L. -25.0 m in Surfers paradise.

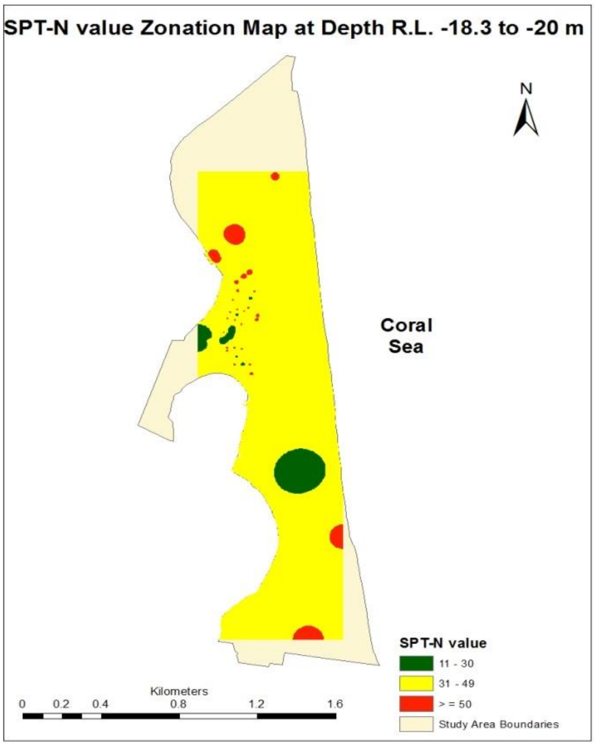


Figure 5. SPT-N zonation map at depth R.L. -18.3 to R.L. -20.0 m

ISBN: 978-1-61208-326-1

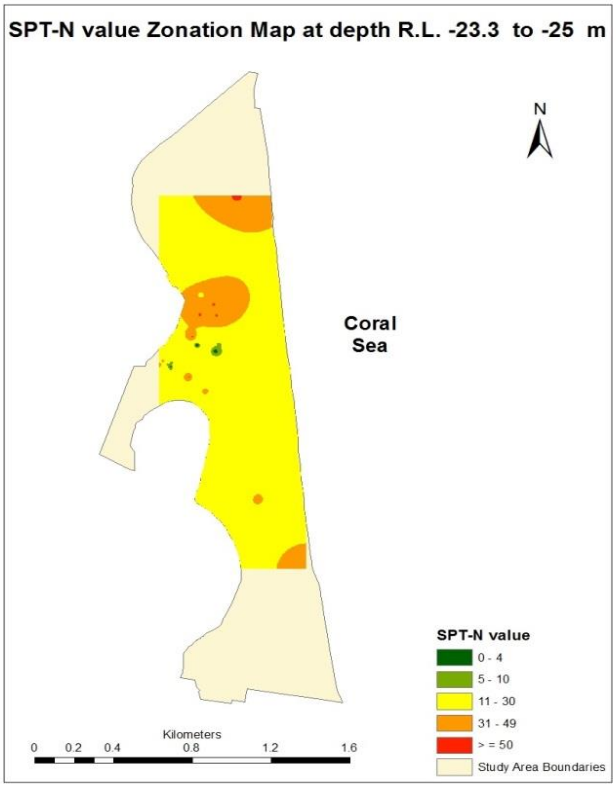


Figure 6. SPT-N zonation map at depth R.L. -23.3 to R.L. -25.0 m

It can be observed from this zonation map (see Fig. 6) that due to the lack of data in some locations at this depth level, only the centre of the Surfers Paradise area has been interpolated. It shows that the medium dense sand and stiff to very stiff clays are dominant with an N value of between 11 - 30 blows. Areas of very soft clays and very loose sand and peat have also been observed.

These zonation maps have shown a representation for the SPT-N value data for all the soil types (sand, clay, peat, and organic clay) in the Surfers Paradise area by using the IDW interpolation technique in Spatial Analyst extension of the ArcMap. This is consistent with what has been reported by Al-Ani et al. [15] and Al-Ani et al. [16]. Similarly, 16 zonation maps have been developed in a densely settled area in Turkey by Orhan and Tosun [4] to determine the suitability of foundation soils. These zonation maps showed three geotechnical properties as follows: 6 zonating the soil class, 6 zonating the SPT-N value, and 4 for the uniaxial compressive strength at 1 m depth intervals from the depth of 1 m and up to the depths of 6.5 m. They also used the IDW technique to accomplish these zonation maps.

Lu and Wong [13] stated that the IDW is one of the most frequently used deterministic models in spatial interpolation. The reason behind this is because it is easy to compute, relatively fast, and straightforward to interpret. In addition, Mueller et al. [17] stated that IDW has been used because it is simple and fast, in contrast to Kriging technique, which is time consuming and more complex than the IDW.

In the literature, there no comparison has been reported among the interpolation methods with the purpose of selecting the suitable method to represent the geotechnical properties of soil in geotechnical engineering. However, a comparison between Ordinary Kriging and the IDW has been performed by [17] to assess soil fertility management in the UK. They showed that there are many opinions regarding the most suitable method in their discipline. Some of them stated that the Ordinary Kriging is better and some stated the IDW is the better option. They suggested doing a cross validation to verify which one is the better method. However, researchers have cautioned using cross validation in the selecting of interpolation methods for mapping purposes [18] and [19]. Further, Mueller et al. [17] stated that cross validation should not be used as the sole criteria for deciding whether or not the specific interpolation method must be used over another.

The Spatial Analyst extension in GIS employs interpolation techniques to create a zonation map. The interpolation, in turn, is a procedure used to predict the values of cells at locations that lack sampled points [20]. As such, the input SPT-N value data have been used to predict SPT-N value in locations that lack sampled points. Therefore, regarding the validation of the results, the resulting zonation maps will be validated with a new data extracted from new engineering projects in the study area. However, from a geotechnical engineering point of view, the interpolated (predicted) values appear sensible as the resulted SPT-N values of the soil in certain selected locations are consistent and comparable with other engineering laboratory tests, such as dry density, moisture content, void ratio, shear strength, and compression index. These physical and engineering properties primarily indicate that there is a valid relationship between, for example, the interpolated high value of the SPT-N value (>50 blows for the very dense sand) and its related soil properties, such as high density (1.92 Mg/m^3), low moisture content (1.5 %), low void ratio (0.23%), low compression index (0.11), and high shear strength (>200 kPa). On the contrary, the resulting SPT-N values which have low SPT-N values (between 0 to 4 blows for the loose sand or organic soil), have magnitudes of low density (0.31 Mg/m^3), high moisture content (239%), high void ratio (4.7 %), high compression index (1.23), and low shear strength (25 kPa). These results are also consistent with what has been reported by Look [14] and Day [21].

Miles and Ho [22] stated that the production of cartographic quality maps or representations can certainly provide support in many pertinent decision making processes. Therefore, this paper provides GIS based zonation maps for the Standard Penetration Test SPT-N value with an aim to facilitate the recognition of the soil stiffness in the study area. In addition, it is type of subsurface visualisation of soil in terms of soil strength at different depth levels. As such, Miles and Ho [22] emphasised that the visualisation can be supplemented by spatial queries of the model results where these queries can help in identifying possible correlations between input parameters and model predictions. Therefore, this paper provides a correlation between the SPT-N values and many geotechnical properties, such as dry density, moisture content, void ratio, compression index, and shear strength.

Further, Player [23] pointed out that GIS can aid geotechnical engineers, as long as the geotechnical data has spatial attributes. Thus, the resulting GIS-based SPT-N value zonation maps can help geotechnical decision makers through providing information on soil stiffness and the occurrence of problematic peat layer in the study area.

IV. CONCLUSION AND FUTURE WORK

Based on the findings, the IDW interpolation technique showed a better representation for Standard Penetration Test N value to characterise the soil in Surfers Paradise. The IDW method has been identified as a better technique based on fixing the predicted points and comparing it with the original input points. It gives an actual representation for each input data compared with other interpolation techniques. Further, the zonation maps showed that the zonation maps developed by the IDW technique exhibited a homogenous distribution for the SPT-N value. Furthermore, the predicted points in the lack sampled areas showed a consistency in terms of the correlation with other geotechnical properties of soil.

Four SPT-N value zonation maps have been developed at various depths. As such, the subsurface profile of the Surfers Paradise area has been examined and showed the occurrence of sands, clays, and peat with various ranges of stiffness magnitudes. This paper is a part of ongoing research to characterise physical and engineering properties of soil in Surfers Paradise, Australia. The next step of this research will be a cross validation for the resulting zonation maps and expand the study area to include the entire city of the Gold Coast, Australia.

V. REFERENCES

[1] C. L. Ho and J. Skeels, ‘GIS modeling of the subsurface stratigraphy’ Proc. 12th Panamerican Conference on Soil Mechanics and Geotechnical Engineering, Cambridge, Massachusetts, USA, 2003, PP. 22-26.

[2] A. Carrara, M. Cardinali, R. Detti, F. Guzzetti, V. Pasqui, and P. Reichenbach ‘GIS techniques and statistical models in evaluating landslide hazard’, Earth Surface and Landforms, vol. 16, no. 5, 1991, pp. 427-45.

[3] M. Xie, T. Esaki, and M. Cai, ‘GIS-based implementation of three-dimensional limit equilibrium approach of slope stability’ Journal of Geotechnical and Geoenvironmental Engineering, vol. 132, no. 5, 2006, pp. 656-660.

[4] A. Orhan and H. Tosun, ‘Visualization of geotechnical data by means of geographic information system: a case study in Eskisehir city (NW Turkey)” Environmental Earth Science, vol. 61, 2010, pp. 455-65.

[5] P. Atkinson and R. Massari, ‘Generalized linear modelling of susceptibility to landsliding in the central Apennines, Italy’ Computer and Geosciences, Vol.24, No.4, 1998, pp.373-385.

[6] R. Higashi and R. Dias, ‘Potentialities of a geotechnical database in a GIS environment of the northern part of Rio Grande Do Sul State-Brazil’, Proc. 12th Panamerican Conference on Soil Mechanics and Geotechnical Engineering, Cambridge, Massachusetts, USA, June 2003, pp. 22-26.

[7] E. Hellawell, J. Lamont-Black, A. Kemp, and J. Hughes, ‘GIS as a toll in geotechnical engineering’, Geotechnical Engineering, vol. 149 (2), 2001, pp. 85-93.

[8] J. Knappett and R. Craig, ‘Craig’s soil mechanics’, 8th ed., London and New York: Spon Press, 2012, pp. 232-233.

[9] J. Bowles, ‘Foundation analysis and design’, 5th ed., New York: McGraw-Hill, 1996, pp.154-156.

[10] C. Lloyd, ‘Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. Journal of Hydrology 308, 2005, pp. 128–150.

[11] J. Ping, C. Green, R. Zartman, and K. Bronson, ‘Exploring spatial dependence of cotton yield using global and local autocorrelation statistics’ Field Crop Research 89 (2–3), 2004, pp. 219–236.

[12] A. Bekele, R. Downer, M. Wolcott, W. Hudnall, and S. Moore, ‘Comparative evaluation of spatial prediction methods in a field experiment for mapping soil potassium’ Soil Science 168 (1), 2003, pp. 15–28.

[13] G. Lu and D. Wong, ‘An adaptive inverse-distance weighting spatial interpolation technique’ Computers & Geosciences, vol. 34, 2008, pp. 1044-55.

[14] B. Look, ‘Handbook of geotechnical investigation and design tables’, London: Taylor & Francis, 2007.

[15] H. Al-Ani, E. Oh, L. Eslami-Andargoli, and G. Chai, ‘Subsurface visualization of peat and soil by using GIS in Surfers Paradise, Southeast Queensland, Australia’, Electronic Journal of Geotechnical Engineering, vol.18, Bund. I, 2013, pp.1761-1774.

[16] H. Al-Ani, L. Eslami-Andargoli, E. Oh, and G. Chai, ‘Categorising geotechnical properties of subsoil in Surfers Paradise using geographic information system (GIS), Proc. 3rd international conference on Geotechnique, Construction Materials and Environment, Nagoya, Japan, November 2013, pp.53-58.

[17] T. Mueller, N. Pusuluri, K. Mathias, P. Cornelius, R. Barnhisel, and S. Shearer, ‘Map quality for ordinary Kriging and inverse distance weighted interpolation’, Soil Science Society of America, 68, 2004, pp. 2042-2047.

[18] N. Cresse, ‘Statistics for spatial data. revised edition’, New York: John Wiley & Sons, 1993.

[19] P. Goovaerts, ‘Geostatistics for natural soil evaluation’, New York: Oxford University Press, 1997.

[20] C. Childs, ‘Interpolating surfaces in ArcGIS spatial analyst’ ArcUser, July-September, 2004, pp. 32-35.

[21] R. Day, ‘Soil testing manual procedures, classification data, and sampling practices’, Ohio: McGraw-Hill, 2001.

[22] S. B. Miles and C. L. Ho, ‘ Applications and issues of GIS as tool for civil engineering modeling’ Journal of Computing in Civil Engineering, vol. 13, 1999, pp. 144-152.

[23] R. S. V. Player, ‘Geographic information system GIS use in geotechnical engineering’ GeoCongress, ASCE, 2006, pp. 1-6.

Comparative Analysis of Photogrammetric Methods for 3D Models for Museums

Unnur Erla Hafstað Ármannsdottir, François Anton, Darka Mioc
Department of Geodesy, National Space Institute, Technical University of Denmark
Elektrovej 328, 2800 Kgs. Lyngby
uhafstad@gmail.com,{fa,mioc}@space.dtu.dk

***Abstract*—The goal of this paper is to make a comparative analysis and selection of methodologies for making 3D models of historical items, buildings and cultural heritage and how to preserve information such as temporary exhibitions and archaeological findings. Two of the methodologies analyzed correspond to 3D models using Sketchup and Designing Reality. Finally, panoramic photography is discussed as a 2D alternative to 3D. Sketchup is a free-ware 3D drawing program and Designing Reality is a commercial program, which uses Structure from motion. For each program/method, the same comparative analysis matrix has been used. Prototypes are made partly or fully and evaluated from the point of view of preservation of information by a museum.**

***Index Terms*—3D Reconstruction, 3D surface models, cylindrical panoramas, Google Sketchup, Designing Reality**

I. INTRODUCTION

There are several museums and institutions around the world that have made, and are in the process of making, 3D models of museum items. According to Mr. Pletinckx [1], the making of 3D models in Cultural Heritage has different purposes: 3D for research, such as the Virtual reconstruction of Regolini-Galassi tomb; 3D for digital restoration, such as bronze disks from the Regolini-Galassi tomb, Italy; 3D to prepare physical restoration, such as Nymphaeum, Sagalassos, Turkey; 3D as documentation, such as Hal Saflieni Hypogeum, Malta; 3D as educational resource, such as Abbey of Saint-John, Biograd, Croatia; or 3D as communication tool, for visual reference and recontextualisation. The virtual reconstruction of the Regolini Galassi tomb, Italy, is a part of the European “Etruscanning 3D" project, which can be followed on the official blog [2]. With the help of 3D modeling, it has been possible to restore a part of the Nymphaeum at Sagalassos, Turkey. All the standing architecture has been conserved and damaged structures, have been repaired using matching materials [3]. Hal Saflieni Hypogeum is an enormous subterranean structure excavated c. 2500 B.C. It is the only known prehistoric underground temple in the world [4]. At the website of Europeana, a 3D pdf can be viewed for further information on this model [5].

3D models of museum items, historical buildings, archeological sites etc. can be used those to create *virtual tours*. At the Virtual Museum Transnational Network(V-MUST) website [6], some of these virtual museums can be visited. In most cases, the viewer can take a virtual tour, making it possible to visit historical places online from anywhere around the world. A virtual tour can be guided or not and can include some or all of the following [6]: written descriptions, photographs and sound-files describing an item or telling a story; maps showing the geographical location of a museum or the internal location of the museum viewer; 3D models visualized as interactive objects and/or videos that make it possible for the viewer to explore an item from multiple view points; hypothesis and historical 3D models that make it possible to tell a story about how a certain place has evolved through centuries, or to visualize an ancient cite that no longer exists.

The goal of this paper is to make a comparative analysis and selection of methodologies for making 3D models of historical items, buildings and cultural heritage and how to preserve information such as temporary exhibitions and archaeological findings. Two of the analyzed methodologies correspond to 3D models using Sketchup [7] and Designing Reality [8]. Finally, panoramic photography is discussed as a 2D alternative to 3D. Sketchup is a free-ware 3D drawing program and Designing Reality is a commercial program, which uses Structure from motion. For each program/method, functionality, the same comparative analysis matrix has been used. Prototypes are made partly or fully and evaluated from the point of view of preservation of information by a museum: the museum of *Byggðasafnið Hvoll*. Due to length, time and museum access constraints, we have not been able to combine these different methods. However, it is unlikely that the results of any linear combination of these methods would be very different than the linear combination of these results.

This paper is organized as follows. In Section II, we present the methodology used in this research, while in Section III, we show the results of our comparative analysis. Finally, we conclude this paper with Section IV.

II. METHODOLOGY

In this section, we will present the different methods used in this paper: “Structure from Motion" using Designing Reality in Section II-A, Sketchup in Section II-B and the cylindrical panorama in Section II-C.

A. Structure from Motion with Designing Reality

`Designing Reality` is a software solution creating 3D models from 2D images. The program is commercial so access to the details of the implementation is not available. However, some examples of 3D models made with the program can be viewed on the official website of Designing Reality: people,

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

cityscapes, statues, landscapes, buildings. According to its founder, Ólafur Haraldsson, the method is less time consuming than normal photogrammetry and, on the official website, it is referred to as *automatic photogrammetry* [8]. Here, it will be referred to as *Structure from motion*.

Structure from motion (SFM) is the process of estimating 3D information from a sequence of 2D images. When the camera moves around an object or the object moves around the camera, information is obtained from images sensed over time. The correspondence between images and the reconstruction of 3D object needs to be found [9]. Once the 2D projection of a point in the real scene has been found, its position in 3D can be assumed somewhere along the ray connecting the camera optical centre and the corresponding spot in the image plane. Tracking its projections across multiple images and using triangulation allows the relatively accurate localisation of the point in 3D [10].

Given the sufficient number of points and lines over images, from different directions it is possible to estimate the 3D location of the structure and the camera location for each point in a 2D image. The 3D structure model differs from the true model by a projective transformation. The process of SFM aims at minimizing the distances between estimated 3D structure projections and actual image measurements [10]. The youtube video “The Structure from Motion Pipeline" [10] explains the following procedure in a very simple way:

1) Use a tripod as object support
2) Add a panoramic head that allows click-stop rotations (Manfrotto 300N).
3) Add object table on top.
4) Position object aligned with the rotation center (see Figure 1).
5) Position lights, and camera on tripod (see Figure 2).
6) Shoot object, rotate one click-stop, shoot again for 360 degrees (A rotation of 10 degrees is recommended here).
7) Shoot another round with different camera inclination if necessary.
8) Remove white background with `Photoshop` or other similar software, replace by fixed color, vignete or image.
9) Do for all images with same parameters.

For reflective objects, use an object tent with surrounding light. Put the object in the tent and install camera through the zipper opening. Then, process as recorded without tent.

As the user has uploaded the images, the procedure of the program is as follows:

- Calculates the location of the camera
- Creates a Point Cloud
- Creates a Mesh (Depth map)
- Adds texture to the Mesh
- Creates an output (video as an option)

B. Drawing of 3D models with Sketchup

`Sketchup` is a 3D modeling tool that allows the user to draw 3D models. The program is available as a free version as well a as professional and due to its simplicity, a user friendly tool-bar and the help of on-line tutorials, almost anybody can use it to develop interesting 3D models [7].

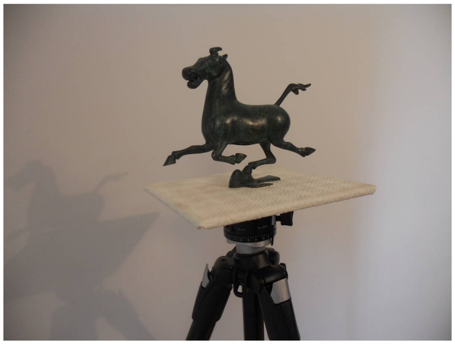


Figure 1. Position object aligned with the rotation center.

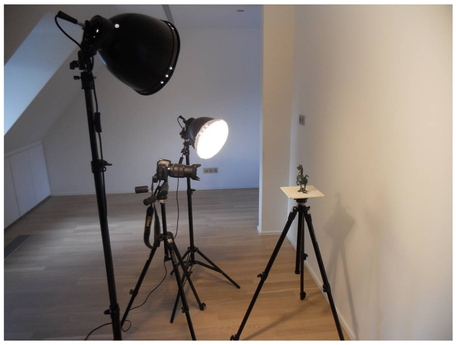


Figure 2. Position lights, and camera on tripod.

In order to investigate the possibilities of developing 3D models for the museum *Byggdasafnid Hvoll*, an introductory survey of surface classes that are easy to build, based on the book Architectural Geometry [11] and some examples of how these classes can be handled with `Sketchup` has been made. The surface classes presented are *rotational surfaces*, *translational surfaces*, *ruled surfaces*, *developables*, *helical surfaces*, *pipe surfaces* and *offsets*. Mathematical description, based on same book, is also shown. Note that some surfaces can be expressed as a surface from more than one surface class. *Rotational surfaces*, sometimes called *surfaces of revolution*, are generated by rotating a plane curve γ, called the profile curve, about a straight line in the plane. If we describe γ as $(f(u), 0, g(u))$, the mathematical expression for rotational surface is $\boldsymbol{\sigma}(u, v) = (f(u)\cos(v), f(u)\sin(v)), g(u))$. Spheres, cylinders, cones and tori are well known rotational surfaces.

If we consider two curves k and l, that intersect at a single point we call the origin o, a *translational surface* can be created by translating one of the curves along the other. Thus, the surface contains a set of curves k_p that are congruent with the profile curve k. The curves are called profile curve and path curve and the same translational surface is generated when changing the roles of these two. Mathematically, each point of the translational surface can be expressed as

$\boldsymbol{\sigma}(u,v) = \mathbf{k}(u)+\mathbf{l}(v)$. The generation of translational surfaces is straightforward and, as they carry two sets of congruent parameter curves, they are commonly used in architecture.

Surfaces that are generated by moving a straight line are called *ruled surfaces*. Cylinders, cones, one-sheet hyperboloids and hyperbolic paraboloids can all be described as ruled surfaces. They carry families of straight lines that are called *generators* or *rulings* and can be described mathematically as $\boldsymbol{\sigma}(u,v) = \boldsymbol{\alpha}(u) + v \cdot \boldsymbol{\omega}(u)$, where $\boldsymbol{\alpha}$ is called the directrix curve. The rulings go through $\boldsymbol{\alpha}$ and are parallel to $\boldsymbol{\omega}$. Ruled surfaces can also be generated by connecting corresponding points of two generating curves and this method gives a wide possibility for generating different surfaces.

Developable surfaces behave just like paper. They can be bent and twisted and unfolded into the plane without stretching or tearing. Due to this they can be easily covered with sheet metal and are therefore widely used in architecture. In addition to this, they carry a family of straight lines which simplifies their construction. Cylinders, cones and tangent surfaces of space curves are all developable ruled surfaces and have the property that a tangent plane is always tangent to the surface along an entire ruling.

Helical surface is created by applying a smooth helical motion to a spatial curve, c. One can generate the same surface either by applying the helical motion to a meridian curve or a cross section. When using the meridian curve $c = (f(v), 0, g(v)$, the mathematical expression for the helical surface is $\boldsymbol{\omega}(u,v) = (f(v)\cos u, f(v)\sin u, g(v) + a \cdot v)$.

A *pipe surface* is the envelope of spheres of equal radius r, whose centers lie on a spine curve c. As an example of two simple pipe surfaces the cylinder has a straight line as the spine curve, and the torus which has a circle as a spine curve. Obviously, non-bendable materials are not suitable for this type of surface. In practice, one uses metal tubes bent into the required form and the manufacturing of such bent tubes is challenging.

In this subsection, two methods for modeling a building are presented. First, the building is modeled from scratch, using non-digital data, that is a drawing on a paper. The interior walls are included in this model. This model could be useful for showing section planes and cuts, or to show the interiors of the house, furniture, etc. Secondly, the same building is modeled using photographs, and here, the model can be exported to Google Earth. Other methods are available, for example by importing a cad file or a drawing and use it as base-plan for the model, and photo-match, those methods are are not explained here.

1) Modeling From Scratch: Here, a three stories house is modeled by using printed plans provided by a realtor. The model contains only straight lines, and the tools that have been used for drawing are *Line*, *Rectangle* and *Push/Pull*. The door and window openings are created with a rectangle that is pushed trough the walls in order to leave an opening. Furnitures are not included. Figure 3 and Figure 4 show the house in two different modes: Shaded with Textures and X-Ray.

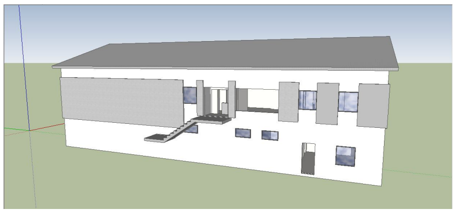


Figure 3. The house shaded with textures.

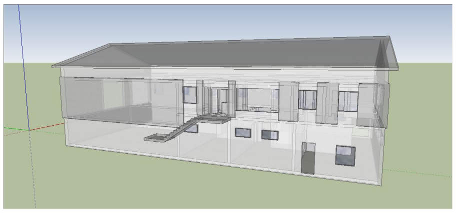


Figure 4. The house in X-ray mode. The interior walls can bee seen.

2) Modeling From Photographs: Google Earth is a useful tool when it comes to presenting 3D models and Sketchup is the most commonly used tool for Google Earth Modeling [12]. There are a few simple steps to follow, and here, the article How to Make a Google Earth Building in SketchUp is used as a guide along with the accompanied video tutorial. At the official Sketchup Help webpage, further guides in Geo-modeling can be found [13].

C. How to Create a Cylindrical Panorama

A cylindrical panorama can be created by shooting several pictures, where each pair of adjacent pictures have an overlap, and then, stitching them together. It is possible to create a cylindrical panorama using any hand-held camera, with any lens, but in order to get good quality, some extra equipment is needed: tripod, spirit level, remote shutter release, Panoramic head ("pano-head") and wide angle lens.

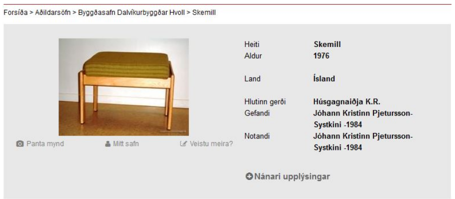
By using a tripod, the camera rotates around a fixed point and at the same level. It also eliminates the risk of camera shake and position change in between shots. If the photographs are not taken at a good level, problems can occur when stitching them together. It might be necessary to crop down the image and thereby reduce the vertical filed of the final result. A spirit level is useful to make sure that the tripods head is correctly leveled, this is especially needed when shooting pictures in the nature. A Remote shutter release can be useful when shooting several photos at a time and also reduces the risk of accidentally moving the tripod while shooting. A Panoramic head is an additional attachment to put on top of the tripod before the camera is attached. It makes it possible to attach the camera such that it rotates around the no-parallax-point of the camera lens. This is important in order to avoid stitching errors. There is as well the feature of click-stops, that allow you to fix the angle of each rotation of the camera and/or a built in spirit level. Using a wide angle lens has the advantage over a regular lens, that less pictures needs to be taken and it is possible to capture more vertically. When choosing a lens it is necessary to make sure that it is supported by the software to be used.

III. Results

In this section, we will see the results of the methods described in last section applied to a museum item (Section III-A) and to a room (Section III-B) and the comparison between different methods (Section III-C)

A. A Museum Item

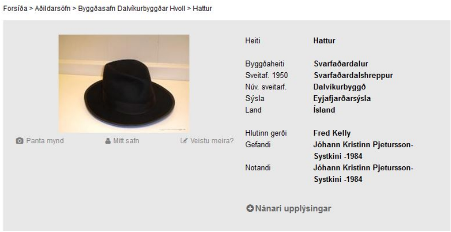
In the database Sarpur [14], the following items from *Jóhannsstofa* are listed with a photo and short description: a custom made footrest, see Figure 5; and a custom made hat, see Figure 6.



Forsíða > Aðildarsöfn > Byggðasafn Dalvíkurbyggðar Hvoll > Skemill

Heiti	Skemill
Aldur	1976
Land	Ísland
Hlutinn gerði	Húsgagnaiðja K.R.
Gefandi	Jóhann Kristinn Pjetursson-Systkini -1984
Notandi	Jóhann Kristinn Pjetursson-Systkini -1984

Figure 5. The footrest as shown in the database Sarpur.



Forsíða > Aðildarsöfn > Byggðasafn Dalvíkurbyggðar Hvoll > Hattur

Heiti	Hattur
Byggðaheiti	Svarfaðardalur
Sveitaf. 1950	Svarfaðardalshreppur
Núv. sveitarf.	Dalvíkurbyggð
Sýsla	Eyjafjarðarsýsla
Land	Ísland
Hlutinn gerði	Fred Kelly
Gefandi	Jóhann Kristinn Pjetursson-Systkini -1984
Notandi	Jóhann Kristinn Pjetursson-Systkini -1984

Figure 6. The hat as shown in the database Sarpur.

1) A Sketchup Model of the Footrest: In this section, a `Sketchup` model of a footrest has been made. The footrest is remarkable due to the extraordinary size and with `Sketchup` it is possible to emphasize this. Figure 7, shows the footrest with scale, and Figure 8 shows the same footrest next to an identical copy, scaled down to regular size, and a 3D woman uploaded from Sketchup warehouse.

2) A Designing Reality Model of the Hat: In this section, a 3D model of a museum object, in this case a hat from *Jóhannsstofa*, is created with photography and the program `Designing Reality`. Völundur Jónsson, photographer, offered his professional help and equipment, taking care of the photos and photo-editing, and Ólafur Haraldsson, the developer of `Designing Reality` created the model using his software. Figure 9 shows a snapshot of the video output of the 3D model. The video can be viewed at vimeo.com [15].

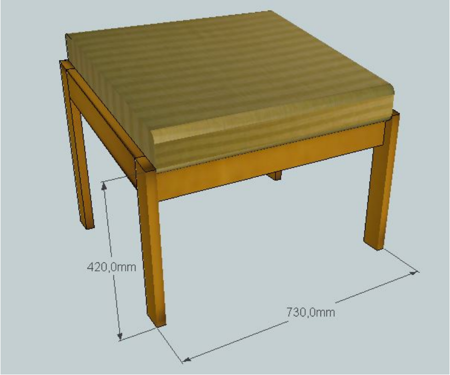


Figure 7. A scaled `Sketchup` model of the footrest.

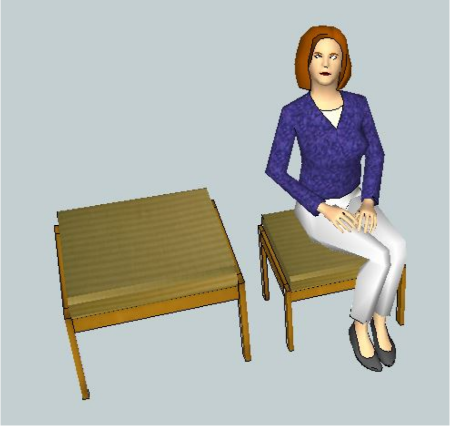


Figure 8. A `Sketchup` model of the footrest, compared to a regularly sized footrest.

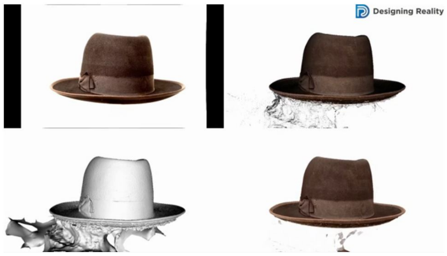


Figure 9. A Designing Reality model of the hat: Photograph, Point cloud, mesh and mesh with texture.

B. A Room

In this section, two methods have been used to partly model and visualize the room *Jóhannsstofa*. In Section III-B1, the room is modeled with `Sketchup` and in Section III-B2, some examples of cylindrical photograps have been made. Völundur Jónsson photographer, kindly offered his professional help and equipment with some of the photos in this section.

1) Modeling the room with Sketchup: In this section, the room is partially modeled with `Sketchup`. Figure 10 shows the model that was made by measuring the room and then, recreating it with `Sketchup`. The radiator was imported

ISBN: 978-1-61208-326-1

from `Sketchup` warehouse and slightly modified. In order to complete the room, texture should be added. Figure 11 shows the same room where the `Sketchup` model of the footrest from section III-A1 has been inserted.

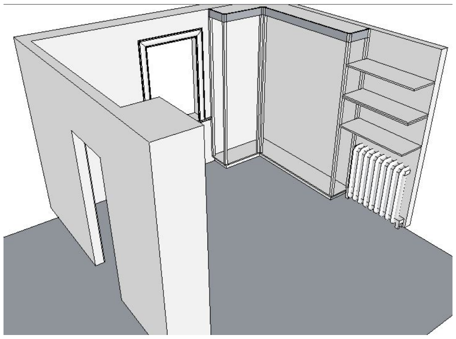


Figure 10. The room partially modeled with Sketchup.

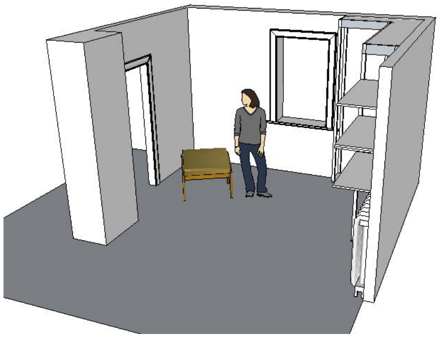


Figure 11. The room partially modeled with Sketchup, and the footrest inserted.

2) Cylindrical Photograph of the Room: In this section experiments have been made with panoramic photographs of the room. Two examples have been created using minimum equipment: only a tripod and a camera, but in each example, with different lenses. Figure 12 shows a panoramic photo of the room. The photo is created with a trial version of Easypano [16] and thus, watermarked. It is taken with a compact digital camera, Canon Ixus 220 HS, and therefore the lens-width is not sufficient for the purpose of creating a nice cylindrical photo showing the room from floor to ceiling. However, since the width of the camera is only 20.8mm, using a tripod and rotating around the center of the camera, the distance from rotation center to the no-parallax-point of the camera lens will not exceed 10.4mm. Therefore, misplacement in the vertical lines is not visible.

In order to get a photo showing a broader view in the vertical, Völundur Jónsson offered his help. Using Canon EOS 5D MKII camera and lens EF 17-40L @ 40mm, f14, a panoramic photo was created that gives the necessary view in order to visualize the museum items from floor to ceiling. Figure 13 by Völundur Jónsson, shows the panoramic picture. The picture is created with the program KOLOR Autopano GIGA 3.0 [17].

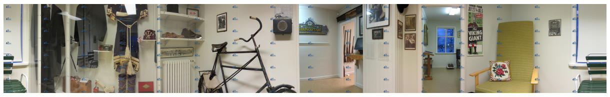


Figure 12. A panoramic photograph of the room, using a handheld camera and a tripod.

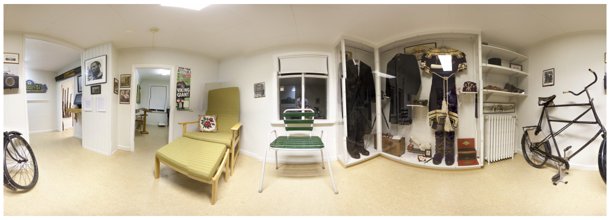


Figure 13. A panoramic photograph of the room, using wide lens and a tripod.

3) Examples of error: When taking a cylindrical photograph there are a few things to consider regarding the photographic equipment. Depending on the materials photographed, some actions might be necessary on the spot. In this case, the glass protecting the museum items should be removed before photographing in order to prevent the mirroring effect shown in Figure 14. This has not been done since the purpose was not to create a virtual tour, but to give examples. Figure 14 shows the vertical misplacement of lines that occurs if the focal point of the camera is not centered.

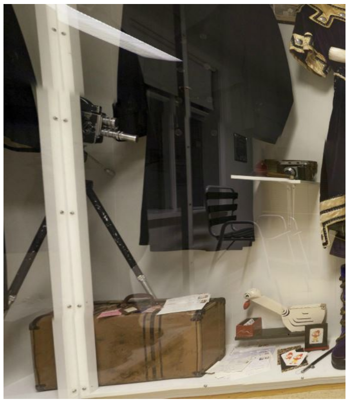


Figure 14. A mirror reflection is clearly seen due to the glass protecting the museum items. Vertical misplacement of lines can be seen as well.

C. Comparative analysis results

At the beginning of the project a personal interview with Íris Ólöf Sigurjónsdóttir was made in September 2012 [18] in order to detect the expectations of the museum. Another interview was made in August 2013 [19], in order to find out how the methods presented fit with those expectations. The following features are considered important to the museum:

cost (as it is a museum with tight budget), user-friendlynes (since there are no IT specialists working for the museum) and suitability for visualizing different items and accuracy. The features above are compared in Table I. `Sketchup` is a freeware, but there is a cost in hiring the staff to implement models. The accuracy is within millimeters, and depending on the modeler, is acceptable when it comes to drawing buildings. However, modeling complex or irregularly shaped items, such as a hat, is not considered to have accaptable accuracy. `Designing Reality` is not a freeware, however, there is less cost in hiring the staff. The accuracy is within millimeters, and is acceptable for complex items. Cylindrical Photography is a 2D alternative, here, not a freeware, and not much training for the staff. The accuracy should not be compared with the other programs since it is not 3D modeling. However, it often shows what needs to be shown. All three methods are considered to be user friendly.

Table I
COMPARISON OF THE METHODS.

Features	Sketchup	Designing Reality	Cylindric. Photogr.
Free	Yes	No	Yes/No
User friendly	Yes	Yes	Yes
vis. existing buildings	Yes	Yes	Yes
vis. non-exist. build.	Yes	No	No
vis. landscape	No	Yes	Yes
complex items	No	3D	2D

IV. CONCLUSIONS

The advantage of using `Sketchup` is mainly the cost, it is a freeware, but also the fact that it is a very user friendly program and compatible with many other programs, such as `Google Earth`. It has been shown that `Sketchup` can handle the surfaces that need to be dealt with in order to create 3D models of the houses. The available data at *Hvoll* is mainly old photographs and drawings and from this data, it is possible to recreate a lost building or a certain room. Furthermore, since many of the houses still exist, it is possible to measure and photograph those. In addition, if the houses have been reconstructed recently, some cad-files exist as well. It is therefore concluded that developing 3D models of historical buildings, using `Sketchup` is a possible option. The disadvantage is that, when it comes to modeling complex items, it can be inaccurate or even impossible to use Sketchup or time consuming, and therefore costly when considering the use of staff-time.

The advantage of using `Designing Reality` is that it is a simple and fast method that gives accurate results. It has been explained how to model an item, but in addition, houses and landscapes can be modeled with the help of drones, helicopters or airplanes. The disadvantage of this program is that it is not possible to make models of non-existing buildings unless the appropriate photographs exist, and the cost is high in this case. However, when compared to `Sketchup` and depending on the size of the model, the cost might be lower considering the staff-time used for each model.

The advantage of taking cylindrical photographs is that it is an easy and fast procedure, and even if not in 3D, it still gives a very clear overview of for example a room or a street-view. It can be used for making virtual tours where already existing materials such as sound-files, videos or 3D models can be included. The disadvantage is that it is usually not a freeware. However using panoramas for the purpose of creating a virtual tour could be thought of as combining it all together, rather than a independent method to compare against those previously mentioned.

The method chosen to model a historical building depends not only on available data, but also on the purpose of the model. There are a few possibilities. Museums could make use of both programs in order to create various forms of teaching material, such as pdf's, slide-shows, animations etc. Alternatively, the purpose could be simply to preserve information about a certain periodical street-view. Furthermore models can be visualized using 3D printers, 3D pdf's or video presentations (animation). Finally, the 3D models created could be used as a part of a virtual tour through the museum.

REFERENCES

[1] D. Pletinckx, "3d as innovation in cultural heritage," http://www.slideshare.net/DanielPletinckx, January 2013, [accessed: 20/11/2013].

[2] "Virtual reconstruction of regolini-galassi tomb," http://regolinigalassi.wordpress.com/, 2013, [accessed: 20/11/2013].

[3] "Sagalassos archaeologocal research project - anastylosis and conservation," http://www.sagalassos.be/en/conservation/anastylosis, [accessed: 20/11/2013].

[4] "Hypogeum of hal-saflieni," http://heritagemalta.org/museums-sites/hal-saflieni-hypogeum/, from Wikipedia, the free encyclopedia.

[5] "The abbey of saint john evangelist in biograd na moru, croatia," http://carare.eu/gre/Media/Files/St-John-Abbey-Biograd-Croatia, cARARE 3D case study.

[6] "v-must virtual museum transnational network," http://v-must.net/, July 2013, [accessed: 20/11/2013].

[7] D. Nath, "History of sketchup," http://www.sketchup-ur-space.com/july11/history-of-sketchup.htm, July 2011, [accessed: 20/11/2013].

[8] "Designing reality," http://www.designingreality.co/, 2012, [accessed: 20/11/2013].

[9] (2004) Structure from motion. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MANESSIS/liter/node10.html. [accessed: 20/11/2013].

[10] (2013, July) Structure from motion pipeline. http://www.youtube.com/watch?v=i7ierVkXYa8. YouTube. [accessed: 20/11/2013].

[11] H. Pottman, A. Asperl, M. Hofer, and A. Kilian", *Architectural Geometry*, D. Bentley, Ed. Bentley Institute Press, 2007.

[12] "How to make a google earth building in sketchup," http://www.staged.com/video?v=ys9b, September 2012, [accessed: 20/11/2013].

[13] "Learning sketchup (a trimble product)," http://support.google.com/sketchup/?hl=en, October 2012, [accessed: 20/11/2013].

[14] "Sarpur," http://www.sarpur.is/, May 2013, [accessed: 20/11/2013].

[15] O. Haraldsson, "Olafur haraldsson," http://vimeo.com/olihar, [accessed: 20/11/2013].

[16] "A trial version of easypano panoweaver," http://www.easypano.com, 2014, [accessed: 20/11/2013].

[17] "Kolor autopano giga 3.0," http://www.kolor.com/image-stitching-software-autopano-giga.html, 2013, [accessed: 20/11/2013].

[18] I. O. Sigurjonsdottir, "Personal interview," September 2012.

[19] ——, "Personal interview," August 2013.