GEOProcessing 2015

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

ISBN: 978-1-61208-383-4

February 22 - 27, 2015

Lisbon, Portugal

GEOProcessing 2015 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 2015

Forward

The seventh edition of The International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2015), held in Lisbon, Portugal, February 22 - 27, 2015, addressed the aspects of managing geographical information and web services.

The goal of the GEOProcessing 2015 conference was to bring 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

GEOProcessing 2015 provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The topics covered aspects from fundamentals to more specialized topics such as 2D & 3D information visualization, web services and geospatial systems, geoinformation processing, and spatial data infrastructure.

We take this opportunity to thank all the members of the GEOProcessing 2015 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 the GEOProcessing 2015. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the GEOProcessing 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that GEOProcessing 2015 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in geographic information research.

We also hope that Lisbon provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

GEOProcessing 2015 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

Yerach Doytsher, Technion - Israel Institute of Technology, Haifa, Israel

GEOProcessing 2015

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
Yerach Doytsher, Technion - Israel Institute of Technology, Haifa, Israel

GEOProcessing 2015 Technical Program Committee

Diana F. Adamatti, Universidade Federal do Rio Grande, Brazil
Ablimit Aji, Emory University, USA
Nuhcan Akçit, Middle East Technical University, Turkey
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
Michela Bertolotto, University College Dublin, Ireland
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
Giuseppe Borruso, University of Trieste, Italy
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
Yao-Yi Chiang, University of Southern California, 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
Joao Ricardo de Freitas Oliveira, INPE - National Institute of Space Research, Brazil
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
Javier Estornell Cremades, Universidad Politecnica de Valencia, Spain
René Estrella, Department of Earth and Environmental Sciences (EES), Katholieke Universiteit Leuven, Belgium
Aly A. Farag, University of Louisville, USA
Nazli Farajidavar, University of Surrey, UK
Marin Ferecatu, Conservatoire national des arts et métiers, France
Lars Fischer, Research Group for IT Security, University of Siegen, Germany
Stewart Fotheringham, University of St Andrews in Scotland, UK
W. Randolph Franklin, Rensselaer Polytechnic Institute - Troy NY, USA
Betsy George, Oracle America Inc., USA
Diego Gonzalez Aguilera, University of Salamanca - Avila, Spain
Björn Gottfried, University of Bremen, Germany
Enguerran Grandchamp, Université des Antilles et de la Guyane, Guadeloupe
Carlos Granell, European Commission - Joint Research Centre, Italy
Malgorzata Hanzl, Technical University of Lodz, Poland
Ahmed AbdelHalim M Hassan, University of Cairo, Egypt
Shuang He, Ecole Centrale de Nantes, France
Erik Hoel, Environmental Systems Research Institute, USA
Zhou Huang, Peking University - Beijing, China
Cengizhan İpbüker, Istanbul Technical University, Turkey
Shuanggen Jin, Shanghai Astronomical Observatory, China
Vana Kalogeraki, Athens University of Economics and Business, Greece
Ibrahim Kamel, University of Sharjah UAE / Concordia University, Canada
Mikhail Kanevski, University of Lausanne, Switzerland
Izabela Karsznia, University of Warsaw, Poland
Rajasekar Karthik, Geographic Information Science and Technology Group, Oak Ridge National Laboratory, USA
Baris Kazar, Oracle America Inc., USA
Margarita Kokla, National Technical University of Athens, Greece
Herbert Kuchen, Westfälische Wilhelms-Universität Münster, Germany
Rosa Lasaponara, CNR, Italy
Robert Laurini, INSA de Lyon - Villeurbanne, France
Ahmed Lbath, Université Grenoble Alpes, France
Dan Lee, Esri, USA
Lassi Lehto, Finnish Geodetic Institute, Finland
Özgür Lütfü Özcep, Hamburg University of Technology, Germany
Fabio Luiz Leite Junior, UEPB - State University of Paraíba, Brazil
Jing Li, University of Denver, USA
Ki-Joune Li, Pusan National University, South Korea
Xun Li, Arizona State University, USA
Jugurta Lisboa, Federal University of Viçosa, Brazil
Qing Liu, CSIRO, Australia
Xuan Liu, IBM T.J. Watson Research Center - Yorktown Heights, USA
Victor Lobo, Portuguese Naval Academy / New University of Lisbon, Portugal
Qifeng Lu, Sevatec Inc., USA
Miguel R. Luaces, University of A Coruña, Spain

Vincenzo (Enzo) Maltese, University of Trento, Italy
Jesus Marti Gavila, Universidad Politecnica de Valencia, Spain
Hervé Martin, Université Joseph Fourier - Grenoble, France
Stephan Mäs, Technische Universität Dresden, Germany
Michael P. McGuire, Towson University, USA
Mark McKenney, Southern Illinois University Edwardsville, USA
Tomas Mildorf, University of West Bohemia - Pilsen, Czech Republic
Beniamino Murgante, University of Basilicata, Italy
Shawn D. Newsam, University of California - Merced, USA
Lena Noack, Royal Observatory of Belgium, Belgium
Daniel Orellana Vintimilla, Charles Darwin Foundation - Galápagos, Ecuador
Özgür L. Özcep, University of Lübeck, Deutschland
Okan Pala, North Carolina State University's Center for Geospatial Analytics, USA
Donna Peuquet, Pennsylvania State University, USA
Maurizio Pollino, ENEA - Italian National Agency for New Technologies - Rome, Italy
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
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, IGIS-NUST Islamabad, Pakistan
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
Paul M. Torrens, University of Maryland - College Park, USA
Luigi Troiano, University of Sannio, Italy
Theodore Tsiligiridis, Agricultural University of Athens, Greece
E. Lynn Usery, U.S. Geological Survey - Rolla, USA
Domitila Violeta Velasco Mansilla, Hydrogeology Group (GHS) Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Barcelona, Spain
Iván Esteban Villalón Turrubiates, Universidad Jesuita de Guadalajara, México
Fusheng Wang, Emory University, USA
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

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Geocoding with OpenStreetMap Data

Konstantin Clemens

Service-centric Networking
Telekom Innovation Laboratories
Technische Universität Berlin, Germany
Email: kclemens@tu-berlin.de

***Abstract*—*OpenStreetMap* (OSM) is a platform where users contribute geographic data. To serve multiple use cases, these data are held in a very generic format. This makes processing and indexing OSM data a challenge. *Nominatim* is an open source search and geocoding engine that consumes OSM data. While Nominatim does process OSM data well, it does not use *term frequency – inversed document frequency* (TF/IDF) based ranking of search results. *Lucene* is a framework offering TF/IDF for ranking of indexed documents. In this paper Nominatim's processing of OSM data is utilized to assemble full addresses with their geocoordinates. These addresses are then indexed in *Elasticsearch*, a web service on top of Lucene. The resulting TF/IDF based geocoding system is benchmarked in comparison with plain Nominatim. The analysis shows: TF/IDF based ranking yields more accurate results, especially for queries with unordered address elements or only partially specified addresses.**

***Keywords*–Geocoding, Address Search, OSM, Nominatim, Elasticsearch**

I. Introduction

OSM data consist of the three entity types *node, way*, and *relation*. For referencing, each entity type has an ID. There are also additional attributes specifying a contributor and a version.

Nodes have values for *longitude* and *latitude*, thus they model points on the globe. Ways compose lines between points by specifying ordered lists of node references. Relations, in turn, may reference both ways and nodes. Therefore relations can model complex geographic features as polygons with holes as well as specify, e.g., a center point for displaying pins on the map. Relations may also reference other relations assembling abstract entities that span several 'real things' such as universities with multiple, wide-spread buildings or groups of islands. Finally, nodes, ways, and relationships can hold an arbitrary number of tags. These key-value pairs specify names, categories, address elements, house number ranges, data sources, speed limits, and other attributes of real-world features that the entities model.

Because of the structure of data, address elements are often spread across different entities. For example, a node might only be tagged with a house number, while the way that references this node only holds the street name information. The way, may be covered by a relation that represents the postal code area. However, the relation not necessarily references the way. Therefore, to offer a geocoding service, addresses need to be assembled out of these OSM entities first.

Nominatim is an open source system that builds on top of OSM data to provide a *geocoding service*. That means, Nominatim resolves named locations (full addresses, or named areas) into latitudes and longitudes of their whereabouts. Nominatim is implemented in multiple programming languages and builds on top of a *PostGIS* enabled *PostgreSQL* data base. It offers a pipeline that parses, assembles, and indexes OSM data. Unlike document stores, Nominatim also precomputes result ranks at indexing time, independent of queries.

In contrast to that, *Lucene* is a generic open source document indexing framework. Lucene supports various ranking schemes for ordering results, including TF/IDF [1][2]. TF/IDF is a formula to rank documents based on query terms and their distributions. Particularly, a document is ranked higher, if it contains query terms that are rarely used in other documents. In the same way, a document is ranked lower, if it only contains query terms that are very common. *Elasticsearch* is a document store that uses Lucene internally to index, find, and rank documents.

In general, geocoding services utilize several algorithms [3][4][5][6]. They have to parse addresses, recognize address elements, derive their meaning, and rank and filter discovered candidates. The index of address data thereby has to be both fuzzy and robust against errors in queries and source data as well as precise enough to rule out ambiguously named address parts that a query did not refer to. TF/IDF based ranking strives to fulfill these requirements with unstructured documents. Because of that, Nominatim is compared to the generic document store Elasticsearch in this paper.

II. Experiment

To compare Nominatim and Elasticsearch, first Nominatim has been set up with OSM data for Europe. Next, the function Nominatim uses to assemble search results has been used to extract all available addresses. Finally, an Elasticsearch instance has been set up next to Nominatim and populated with the extracted addresses. This way two geocoders with the same addresses indexed were running next to each other: Nominatim with pre-ranked results as well as TF/IDF based Elasticsearch. Note that while Nominatim has loaded and indexed all of OSM data, only assembled addresses have been indexed in Elasticsearch. Thus, extended features of OSM and Nominatim, e.g., translated addresses, were not available in Elasticsearch.

For the first experiment, 2000 randomly selected addresses were extracted from the Nominatim database. In accordance to the set up, the addresses were from various European countries and were all indexed in both systems. From these addresses queries with exact addresses have been generated first. The

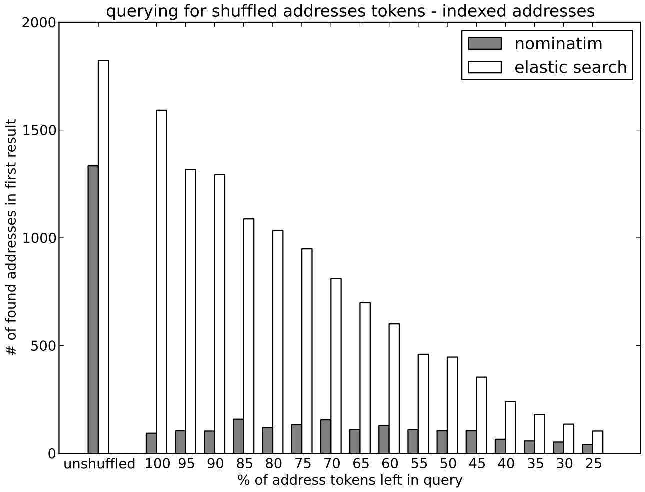


Figure 1. Nominatim and Elasticsearch on full and partial addresses.

next set of queries contained shuffled address tokens. This simulated queries that do not adhere to an address format. Finally in steps of 5% random address tokens have been removed from the queries, simulating incomplete requests. A request was regarded as successfully served, if the first result of the response was the address the query has been generated from. The number of successful requests of the respective systems is displayed in Figure 1.

The second experiment used 1000 addresses of German pharmacies. Only addresses indexed in both systems have been used. To count successful requests, all addresses have been geocoded with Google's geocoder. Because Google data and OSM data differ, different coordinates for same results have been expected. Therefore distances of the first result to Google's geocoordinates have been grouped into buckets. The buckets *within 100m, 100m-1000m, further than 1000m*, and *no result* have been used. For a general purpose geocoder, only the first bucket should be treated as a successful result. Similar to the first experiment, requests with full addresses as well as requests with only 75% of the address tokens have been stated. Table I enumerates the results.

TABLE I.
NOMINATIM AND ELASTICSEARCH ON PHARMACY ADDRESSES.

full addresses	*within 100m*	*100m-1000m*	*further 1000m*	*no result*
Nominatim	946	0	0	54
Elasticsearch	1000	0	0	0
75% address	*within 100m*	*100m-1000m*	*further 1000m*	*no result*
Nominatim	399	13	0	588
Elasticsearch	988	11	1	0

Note that addresses used in the first experiment have been generated by Nominatim. They contained many additional address elements stating administrative areas that are not commonly mentioned in German postal addresses. Pharmacy addresses where of that common type: They only contained the four address elements house number, street, city, and postal code. This is why for the second experiment only addresses with 75% address tokens have been used next to the full addresses: Most of them lacked exactly one token which made up an entire address element. Also some of the chosen pharmacies were already present as POIs in the OSM data set in addition to their addresses. Still, because the pharmacies' addresses have been used to create requests, the presence or absence of a POI in OSM has not influenced the result.

III. RESULTS

Figure 1 shows that Nominatim is incapable to deal with shuffled address tokens: it recognizes only a small fraction of the addresses that Elasticsearch can find. While more missing address tokens lead to less accurate results of Elasticsearch proportionally, the performance of Nominatim at first increases slightly. Because less tokens can be shuffled in less ways, Nominatim reaches it's sweet spot at ca. 70% of address tokens being queried. After that addresses become less distinctive and Nominatim decreases in accuracy in the same way as Elasticsearch. Also, already for full addresses Elasticsearch outperforms Nominatim.

According to Table I, Nominatim fails to geocode 54 full addresses. All full addresses are resolved by Elasticsearch to 100m correctly. Again, there is a big drop in Nominatim's performance for incomplete addresses with shuffled tokens – only 399 addresses are still found, while Elasticsearch manages to geocode 988 addresses into the *within 100m* bucket.

IV. CONCLUSION

The experiments showed that Nominatim is tightly coupled to specific address formats. This is a strong limitation, as there are contradicting address formats world wide [7]. It is also clear that preranking results independent of queries leads to less accurate results. It is worth noting that the experiments only took queries for street addresses with house numbers into account. The actual ratio of queries for named areas should be incorporated into benchmarks of live services.

Elasticsearch has proven to be more robust against shuffled and dropped address tokens. Because of that, for geocoding addresses Elasticsearch populated with addresses assembled by Nominatim yields better results than Nominatim alone. It is also clear that TF/IDF ranking is more suitable for geocoding than precomputing result ranks. Obviously, the actual performance of a geocoding system highly depends on the actual queries. Still, Elasticsearch can be used as a solid base line when developing and comparing geocoding algorithms and indexes.

REFERENCES

[1] G. Salton and C.-S. Yang, "On the specification of term values in automatic indexing," Journal of documentation, vol. 29, no. 4, 1973, pp. 351–372.

[2] G. Salton, C.-S. Yang, and C. T. Yu, "A theory of term importance in automatic text analysis," Journal of the American society for Information Science, vol. 26, no. 1, 1975, pp. 33–44.

[3] J. Fitzke and R. Atkinson, "Ogc best practices document: Gazetteer service-application profile of the web feature service implementation specification-0.9. 3," Open Geospatial Consortium, 2006.

[4] D. Goldberg, J. Wilson, and C. Knoblock, "From text to geographic coordinates: The current state of geocoding," URISA-WASHINGTON DC-, vol. 19, no. 1, 2007, p. 33.

[5] D. Yang, L. Bilaver, O. Hayes, and R. Goerge, "Improving geocoding practices: evaluation of geocoding tools," Journal of Medical Systems, vol. 28, no. 4, 2004, pp. 361–370.

[6] L. Can, Z. Qian, M. Xiaofeng, and L. Wenyin, "Postal address detection from web documents," in Web Information Retrieval and Integration, 2005. WIRI'05. Proceedings. International Workshop on Challenges in. IEEE, 2005, pp. 40–45.

[7] K. Clemens, "Automated processing of postal addresses," in GEOProcessing 2013, The Fifth International Conference on Advanced Geographic Information Systems, Applications, and Services, 2013, pp. 155–160.

Using Geosynchronization for Incremental Update of INSPIRE Service Databases

Lassi Lehto, Eero Hietanen and Pekka Latvala
Department of Geoinformatics and Cartography
Finnish Geodetic Institute
Masala, Finland
lassi.lehto@fgi.fi, eero.hietanen@fgi.fi, pekka.latvala@fgi.fi

***Abstract*—Opening of governmental geospatial datasets has lead to the creation of multitude of copies of these datasets, used as source information for various value adding services. Obligations set forth by the INSPIRE directive bring new data resources available via service databases, copied from the original sources. Without proper updating procedure these copy datasets become fast out-of-date. The Geosynchronization Service (GSS) proposal of the Open Geospatial Consortium (OGC) is a promising approach for resolving this problem. As an example, GSS is investigated in this paper as a tool for synchronizing INSPIRE datasets and keeping service databases up-to-date in a multi-provider situation.**

Keywords-geosynchronization; incremental update; service database; open data; INSPIRE.

I. INTRODUCTION

Opening of governmental data resources is one of the most important trends affecting the provision of geospatial services in Europe. As datasets become freely available, many different actors, also in private sector, are able to simply download a dataset, combine it with some application-specific content or another open public sector dataset and create a dedicated online customer service that may have thousands of active users.

Without proper handling of updates the service databases may become quite fast out-of-date. In some cases the governmental datasets are being updated on a weekly or even daily basis. A copied service database can often be left untouched for a long time. This can potentially create a huge data quality problem.

In the context of the INSPIRE directive, European providers of spatial datasets are obliged to develop standardized services for content themes that in many cases involve multiple separate institutions [1][2]. In this situation, it is usually necessary to set up separate service databases, in which the content is integrated, harmonized and brought to the specified common schema. Again, the problem of data decay over time is obvious. The services become fast out-of-date, unless the service databases are regularly updated from the original sources.

A full data download and a complete replacement of the service database contents is often unpractical, due to vast data volumes, or local additions made in the copy dataset. Thus partial, feature level update of the service database becomes a necessity. The principles of incremental update have been investigated over the years in various research projects. The latest action in this area is the work of Open Geospatial Consortium (OGC) to specify a service type called Geosynchronization Service (GSS) that would allow incremental update of data resources over the network [3]. In this paper, the GSS approach is tested in the context of an INSPIRE service development project, in which the service database is composed of content coming from two separate data providers. The rest of the paper is organized as follows. Section II discusses the past work done in the area of incremental update and synchronization of geospatial data. Section III describes the principles of the OGC's GSS specification. Section IV describes the context of the case study and the implementations developed in it. Section V concludes the paper with discussion on the lessons learnt and prospects for the future work.

II. PREVIOUS WORK

Incremental update has long been studied as a mechanism for synchronizing geospatial datasets across organizational borders [4][5]. Mostly these studies did not consider network-based solutions. Milanović et al. describe an example of an early service-oriented implementation [6]. In this experiment general OGC service standards were used to synchronize datasets among participating data providers. The examples, in which GSS specification has been adopted as a solution for data synchronization include distributed maintenance of the National Hydrography Dataset in Indiana [7] and New Hampshire [8]. A few software implementations have been developed based on the GSS specification, like CarbonCloud Sync of Carbon Project Inc. [9], userSmarts of ImageMatters [10] and GeoGig of Boundless [11].

III. OGC'S GEOSYNCHRONIZATION SERVICE

A. *Principles*

The Geosynchronization specification of OGC is based on the idea of collaborative capture of geographic information, in which one party acts as a Publisher proposing changes to be made to the target database [12]. Another important role is that of a Reviewer, who checks the proposed changes and either accepts or rejects them, based on some well-established quality criteria. The third role is a Subscriber, a party that is interested in receiving the authorized changes made to the target database to keep a copied dataset up-to-date, i.e. synchronized with the original.

These functionalities are formulated as a set of Web Service interfaces that support the needs of the identified actor roles. The main workflow of the Geosynchronization Service is depicted in the Figure 1.

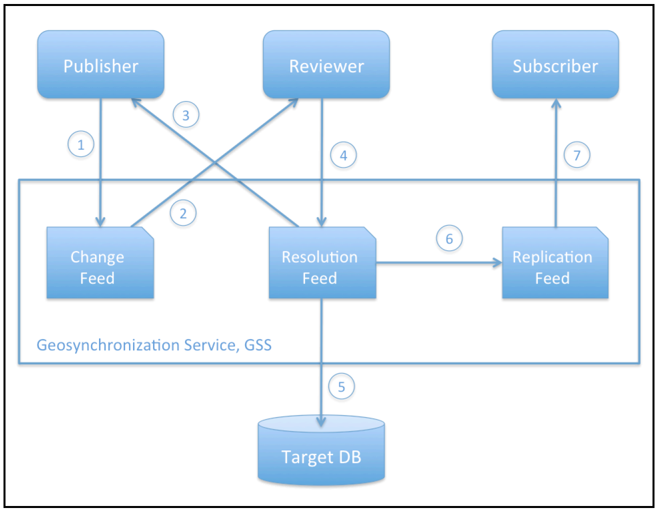


Figure 1. The main Geosynchronization Service workflow.

The Geosynchronization workflow commences with a Publisher proposing a change to be made in the target database (action number 1). This proposal is fed into the ATOM stream called the Change Feed. Looking into the change proposal from the Change Feed the Reviewer makes a decision, whether the change should be accepted or rejected (2). Thus the Reviewer represents the organization having responsibility and authority on the target database. If a given change proposal gets accepted by the Reviewer, information on the decision is sent to the respective Publisher via the Resolution Feed (3), the change is committed to the database (5) and the corresponding record of the change is created to the Replication Feed (6). If the change is rejected, only note of this decision is returned to the Publisher and no other action is taken. Finally, if a third party is interested in the changes happening in the target database, that party can register as a Subscriber to receive the changes from the Replication Feed (7).

All the important information in the Geosynchronization Service is conveyed through standard ATOM feeds [13]. These include Change Feed to manage the proposed updates, Resolution Feed to convey information on the validation decisions, and Replication Feed for synchronizing a remote copy of the dataset with the master database managed within the Geosynchronization Service. This mechanism facilitates a standards-based approach for retrieving only items that have been created since the previous request. The change requests are encoded as Web Feature Service (WFS) Transaction operations [14].

Main use cases of the Geosynchronization Service include maintenance of multi-resolution datasets by organizations on different levels of administration hierarchy (for instance municipality proposing updates on a dataset managed by a state level authority) and crowd-sourced provision of updates to a governmental dataset. In both cases an outside actor (Publisher) proposes changes to be made on a dataset managed by an authority (Reviewer) that has to validate the changes before they are committed to the database. A further third party can then act as a Subscriber to keep a copied instance of the dataset up-to-date by checking the Replication Feed regularly and applying the changes to the copy dataset. The Replication Feed can be queried with full filtering capabilities, so that the Subscriber can only receive changes that are relevant for his purposes. If deemed appropriate, a GSS can be configured to automatically commit all changes directly to the Subscriber’s database. A Subscriber can also opt for receiving only change proposals with information of their validity (Resolution Feed).

B. *Status of the Specification*

The Geosynchronization Service specification has been initially drafted in the context of the OGC Interoperability Program as a Canadian Geospatial Data Infrastructure (CGDI) pilot in 2007 - 2008. Since then the specification has been amended and refined in other OGC and local Canadian programs. The candidate standard was published as OGC Engineering Report in 2011. Since 2011 the GSS specification has not advanced in the OGC process and the official standard is not yet available.

IV. CASE: INSPIRE SERVICES IN A MULTI-PROVIDER SITUATION

The case implementation of the GSS specification described in this paper has been developed in the context of the European INSPIRE directive and the requirement it sets for EU member states to establish standardized services on geospatial content, as defined in common data specifications and content schemas. In many cases, the content specified in one single INSPIRE theme, like Hydrography (HY), involves datasets maintained by many state level organizations. The commonly agreed European level content schemas also often significantly differ from the data structures used in national datasets. All this necessitates maintenance of separate service databases, in which content is integrated from different data provider organizations and the data structures are designed to support INSPIRE schemas. This environment represents an ideal use case for a GSS piloting. On one hand, the data provider organizations involved need to keep their datasets synchronized, otherwise the quality of the resulting service output would be severely degraded. On the other hand, the service database has to be kept up-to-date by replicating to the service database the changes made in the original datasets.

In the case implementation discussed here, the organizations involved in the provision of INSPIRE Hydrography theme are the National Land Survey of Finland (NLS), the official national mapping authority, and the Finnish Environmental Institute (SYKE), a governmental organization responsible for supporting the state activities related to the protection of environment. The mapping authority maintains an up-to-date geometric representation of the physical hydrologic features and the environmental institute manages a rich record of data values describing the status of those physical features. SYKE also maintains a network representation of the hydrography system, so there is a need to synchronize spatial geometries between the two organizations too.

The test implementation was designed according to the principles of the GSS, predominantly to synchronize the hydrography network geometries managed by SYKE, based on the frequently updated geometries of physical waters, maintained by NLS (see Figure 2). To facilitate the flow of updates from NLS to SYKE, a prototype GSS implementation was developed. As the data updates in this case can be seen as authoritative changes, the review/validation step was not included in the prototype and is thus not further discussed in the paper.

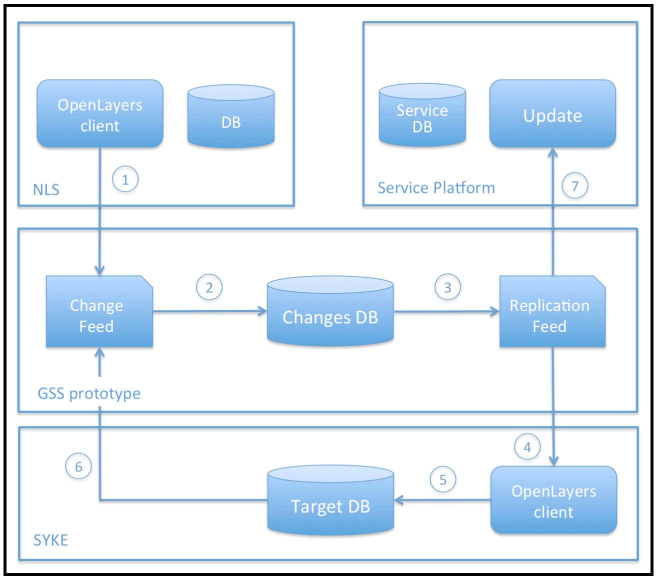


Figure 2. The GSS prototype implementation architecture.

The GSS prototype uses PostGIS/PostgreSQL database (Changes DB) for storing the proposed changes. In the demonstration environment, the changes are produced using OpenLayers-based client application that formulates the changes as WFS-T messages (1). Once the GSS prototype receives a WFS-T message, it parses the message extracting all relevant information, and stores everything to the Changes database (2). This functionality is implemented as a Java Servlet. The Replication Feed is produced by another Java Servlet that, once receiving a replication request, processes it, retrieves the needed pieces of information from the Changes database and then formulates the response in the form of an ATOM feed (3).

The Changes database was designed to be simple. One single table contains all data items related to proposed changes. A record in the table corresponds to one change proposed to an individual feature. The columns of this database table are described in Table I.

The fields stored in the Changes database table are carefully selected to support efficient serving of the Replication Feed queries. The WFS-T DELETE operation represents a special challenge, as it does not contain the attributes of the deleted feature, but only its identification. As the approach requires the geometry of the affected feature before the change to be stored into the change table, this information has to be acquired elsewhere. In the prototype implementation this is done by requesting the feature information from the target database before the change in question is applied to it. To improve storage efficiency, the geometry of the feature in the Transaction column could also be removed and dynamically generated from the Geom column when needed.

TABLE I. COLUMNS OF THE CHANGES DATABASE TABLE.

Column	Data type	Description
GID	integer	Unique identifier of the feature
Type	string	Name of the feature class
ChangeTime	date/time	Timestamp of the change
Operation	string	INSERT/UPDATE/DELETE
Transaction	string	The change, encoded as a WFS-T transaction message
Geom	geometry	Geometry of the feature after the change. If operation is DELETE, then before the change.
EntryID	integer	Unique identifier of the change
Author	string	Publisher of the change

As the Changes database's Transaction column contains a complete representation of the affected feature after the change in every UPDATE operation, the Changes database also serves as a fully versioned data store. All known versions of a given feature can be requested from the GSS Replication Feed using the FEATUREID parameter. If the initial creation of the feature is also recorded as an INSERT operation, then the full history of the feature is available from the GSS.

The Replication Feed can be queried from the service by GetEntries operation. In this query the calling application can limit the result set by the following predicates: STARTTIME/ENDTIME (limits for the change timestamp), BBOX/GEOM (spatial extent inside which the changes are requested, as a rectangular bounding box or free form polygon), FEATUREID (returns changes related to the given feature or features), ENTRYID (returns the indicated change). The resulting Replication Feed is formatted as an XML-encoded ATOM feed, following the principles established in the GSS specification.

To facilitate illustrative visualization of the changes in a map display, an extra GetEntryIdObjectGeometry operation has been added to the prototype as an extension for the GSS specification. This enables geometries of an identified change to be requested from the Changes database. As the original change messages are stored in a separate column (Transaction), the Replication Feed can be quite efficiently generated.

Two OpenLayers-based map viewer user interfaces have been developed to demonstrate the functionality of the GSS.

One is used to initiate changes. The source dataset is queried via a WFS interface from the NLS DB and the transactions resulting from the edits done on the features in the map window are redirected to the prototype GSS instance. The other map viewer is used to visualize features from the SYKE's target database that is being synchronized with the original NLS database. By the client application a user can visualize the contents of the target database, together with the changes requested from the Changes database via the GSS instance (4). The user can then interactively select the changes that seem appropriate for the target database. Finally the selected changes are applied to the database through the WFS-T interface (5). The edits done on the SYKE's target database will again trigger change requests in the GSS's Change Feed, relating to the feature types maintained by SYKE (6)

In a similar fashion, the GSS Replication Feed can be accessed to keep a remote service database up-to-date (7). This could be a completely automated process as all the changes are authoritative and can be copied without further consideration to the service database.

V. CONCLUSIONS

The prototype GSS implementation described in the paper has been developed in the context of an INSPIRE Download Service piloting project. The experiences gained in the work point out that the GSS approach can be utilized to synchronize two databases for achieving consistent representation of geospatial datasets and thus for facilitating the integration of INSPIRE content in multi-provider environment. At the same time the GSS serves as a tool for keeping the copied service database up-to-date.

The GSS developed in the project is essentially based on standardized approaches. The update information is transmitted and stored as standard WFS-T transactions. GSS response messages are formulated as ATOM-conformant feeds. OGC-compliant service interfaces are used to access the data stores, both to retrieve and to modify the contents of the backend databases. The selected approach facilitates integration and usage of the proposed system with various off-the-shelf commercial and open source solutions.

An interactive process has been developed for selecting relevant updates using the capable filtering options of the GSS query interface. A map interface-based tool is used for fine-tuning the set of updates to be applied to the target database. These functions enable highly varied application needs to be taken into account in the update process.

The experiences gained in developing the GSS implementation confirm that the approach taken in the OGC's Geosynchronization activity is usable and can be adopted in the context of multi-provider INSPIRE Download Service provision.

Future developments could involve for instance adding of schema transformation capabilities to the GSS to enable synchronization among heterogeneous schema structures. Another need for further investigation is the scalability of the proposed approach in case of larger data volumes and various types of data, and the methods, by which this could be ensured.

ACKNOWLEDGMENT

The work described in this paper has been co-funded by the National Land Survey of Finland, the Finnish Environmental Institute, the Finnish Ministry of Agriculture and Forestry and the Finnish Geodetic Institute.

REFERENCES

[1] European Commission, “INSPIRE Directive", 2007. at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF [accessed: 2014-10-09].

[2] European Commission, “COMMISSION REGULATION (EC) No 976/2009 of 19 October 2009 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards the Network Services”, 2009. at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX: 02009R0976- 20101228&from=EN. [accessed: 2014-10-09].

[3] Open Geospatial Consortium, Specifications Working Group on GeoSynchronization Home Page. at: http://www.opengeospatial.org/projects/groups/geosyncswg [accessed: 2014-10-09].

[4] M. Scheu, W. Effenberg, and I. P. Williamson, “Incremental update and upgrade of spatial data”, 2000. Zeitschrift fur Vermessungswesen, 125(4), 2000, pp. 115-120.

[5] A. Cooper and A. Peled, “Incremental Updating and Versioning”, 2001. 20th International Cartographic Conference, Beijing, Vol 4, 2001, pp. 2804-2809.

[6] A. Milanović, J. Guimet, E. Rodellas, and M. A. Bolívar, “Interorganizational Geo-Synchronization Using Open Geospatial Consortium's (OGC) Technologies to Share and Harmonize Data in Catalonia”, INSPIRE Conference 2010, Krakow, at: http://geoportal.cat/geoportal/eng/documents/articles/Ponencia_INSPIRE_2010.pdf [accessed: 2014-10-13].

[7] J. Sparks, “Indiana High & Local-Resolution NHD Update Geosynchronization”. Framework Data Exchange through Automated Geosynchronization , Final Report, 2011. at: https://www.fgdc.gov/grants/2010CAP/InterimFinalReports/242-10-2-IN-FinalReport.pdf [accessed: 2014-10-13].

[8] R. Chormann and J. Harrison, “New Hampshire NHD GeoSynchronization Network“, 2010 NSDI CAP Final Report, 2012, at: https://www.fgdc.gov/grants/2010CAP/InterimFinalReports/238-10-2-NH-FinalReport.pdf [accessed: 2014-10-10].

[9] N. Goldstein, “GeoSynchronization – Synchronization of Geo-Databases through Interoperability”, 2010. at: http://proceedings.esri.com/library/userconf/feduc10/papers/user/goldstein.pdf [acessed: 2014-10-10].

[10] ImageMatters, userSmarts Home Page, at: http://www.imagemattersllc.com/products/usersmarts.php [accessed: 2014-10-13].

[11] GeoGig Home page, at: http://geogig.org [accessed 2014-10-13].

[12] Open Geospatial Consortium, “OWS 7 Engineering Report – Geosynchronization Service”, 2011. at: http://portal.opengeospatial.org/files/?artifact_id=39476 [accessed: 2014-10-09].

[13] M. Nottingham and R. Sayre [eds.], “The Atom Syndication Format”, 2005. at: http://tools.ietf.org/html/rfc4287 [accessed: 2014-10-09].

[14] P. A. Vretanos [ed.], “OpenGIS Web Feature Service 2.0 Interface Standard”, 2011. at: http://portal.opengeospatial.org/files/?artifact_id=39967 [accessed: 2014-10-09].

PABench: Designing a Taxonomy and Implementing a Benchmark for Spatial Entity Matching

Bilal Berjawi*, Fabien Duchateau†, Franck Favetta‡, Maryvonne Miquel* and Robert Laurini*
*LIRIS, UMR 5205
INSA de Lyon, Lyon, France
†LIRIS, UMR 5205
Université Claude Bernard Lyon 1, Lyon, France
‡LIRIS, UMR 5205
Ecole Nationale Supérieure de la Nature et du Paysage, Blois, France
Email: firstname.lastname@liris.cnrs.fr

Abstract—**The tremendous increase of data sources containing spatial information is bound up with the diversity of geospatial applications such as location-based services (LBS) and global positioning systems. LBS providers use maps to locate spatial entities referring to points of interest (POI), for instance restaurants or locations of events. In our study, we specifically focus on places that tourists can get through LBS. The multiplication of these providers has an impact on the quality of data: spatial entities referring to the same POI may include spatial and terminological properties with incomplete, inconsistent, inaccurate or even wrong data. Thus, entity matching approaches have been proposed to discover correspondences between spatial entities, and experimental validations are traditionally performed to demonstrate the performance of these approaches in terms of effectiveness and efficiency. However, the datasets used in these experiments are rarely available, thus limiting their reuse and a fair comparison between the proposed approaches. This paper introduces PABench, a benchmark for spatial entity matching, which is available for researchers to assess their algorithms. Our benchmark includes a taxonomy of observed differences, inconsistencies and errors, which enables the characterization of different LBS providers. PABench can provide a complete and accurate evaluation of the different aspects of spatial entity matching approaches, and also facilitate an understanding of their weaknesses and abilities with respect to spatial integration. This paper provides a conceptual platform to enable LBS interoperability.**

***Keywords*–Benchmark; Location-based services composition; Location-based services interoperability; Spatial entity matching; Taxonomy.**

I. INTRODUCTION

With the proliferation of LBS and the increasing amount of geographic data, many issues arise related to the integration of several sources of spatial data. This integration is necessary in order to update information that changes daily [1] or to produce more complete and accurate information [2]. LBS are daily used in various applications, and cartographic providers (e.g., Google Maps, Bing Maps, MapQuest) play an essential role in offering POIs such as restaurants, hotels and tourist places. A POI can be defined as a geographic object that has a point geometric shape. A POI has spatial attributes, longitude and latitude, and terminological (non-spatial) attributes such as name and type (e.g., restaurant, hotel). Some providers may supply additional terminological attributes such as address, phone number, web site, customers' ratings, etc. Due to lack of completeness, noise, inaccurate and contradictory data, it is interesting to propose solutions for detecting corresponding entities (i.e., which refer to the same POI) as given from different providers. This challenge aims at improving the quality and the relevance of information, which has a significant impact in tourist applications.

Geospatial integration has been widely studied under the term "*map conflation*" where two whole maps are integrated. Integration of maps consists of identifying the corresponding entities and fusing them [3]. Ruiz et al. present a wide description of the art with respect to map conflation [4]. The authors describe existing works in map conflation regarding their formats (raster and vector) and their criteria such as spatial data, terminological data and topological relationships between entities. Some works have been proposed in map conflation using punctual entities [5][6][7], linear entities [8][9][10] and polygonal entities [11][12][13]. Thakkar et al. propose a tool to assess the quality of geospatial data sources [14]. It utilizes the information from geospatial data sources with known quality to estimate the quality of geospatial data sources with unknown quality. In the last decade, the integration problem mainly refers to the "*entity matching*" research domain, enhanced by a spatial aspect. The discovery of corresponding entities is performed either by exploiting only spatial information [15] or by computing and combining terminological similarities for selected attributes (e.g., name, type) [16][17]. Machine learning algorithms may be applied to tune the parameters (e.g., weights) of a matching process [18]. When corresponding entities have been detected, an interesting use case aims at displaying a merged entity, i.e., to use a crafted algorithm to fuse the attributes' values of these corresponding entities. Such merging algorithms are not 100% confident. For instance, two corresponding entities may have a different location and the algorithm needs to determine the correct position. Similarly, the names or the phone numbers of two corresponding entities may differ, and the choice of the correct values relies on the merging algorithm. A merged entity may therefore include at different levels some uncertainties, which have to be presented to end-users [19].

Unfortunately, it is impossible to have a fair evaluation and comparison between existing approaches due to the absence of benchmarks. Our objective is to design such a benchmarking system, called PABench (POI Alignment Benchmark) to compare the existing spatial entity matching approaches in order to improve them or to build a better one. First, it is important to find out how the corresponding entities differ between each

ISBN: 978-1-61208-383-4

other in order to understand how they can be integrated. To do so, we formalize a taxonomy to distinguish the differences that may occur within the entity set of one single provider, denoted as Intra-Difference class, and the differences between the entity sets of several providers, denoted as Inter-Difference class. The former class helps evaluate the quality of the entity set of one provider (e.g., complete information, redundancies), while the latter helps evaluate the matching between the entity sets of several providers. This taxonomy is useful to produce statistics about the providers' datasets. The next step is to create a benchmark based on the characterization of the taxonomy to understand the weaknesses and strengths of spatial entity matching approaches.

Our contributions can be summarized as follows: (i) propose a model of LBS, (ii) construct the taxonomy and understand how it impacts the results' quality of a spatial entity matching system, (iii) create a benchmark that serves in evaluating and comparing the spatial entity matching approaches [20].

The paper is organized as follows. In Section II, related work is discussed. The taxonomy of differences is given in Section III. PABench is presented in Section IV. The construction of datasets is given in Section V and a use case is provided in Section VI. Finally, Section VII concludes the paper and outlines future work.

II. Related Work

In this section, we present existing benchmarks in data integration.

In **ontology matching**, the objective is to discover semantic correspondences between concepts and properties of different ontologies [21]. Ontology alignment researchers have designed Ontology Alignment Evaluation Initiative (OAEI) [22] to compare ontology alignment tools. The OAEI datasets fulfill various criteria. For instance, the *benchmark* dataset gathers many ontologies in which a specific type of information has been altered (modifications, deletions, etc.). Consequently, it aims at detecting the weaknesses of a tool according to available information. Other datasets might be very specific like the *Food and Agriculture Organization* ontologies, which usually require external resources such as dictionaries to obtain acceptable results. In 2013, the initial datasets have been extended with synthetic ones [23].

Schema matching and mapping can be defined as the discovery of correspondences between schema elements as well as the mapping functions to transform source instances into target instances [24]. The community has designed benchmarks for evaluating these two tasks. XBenchMatch enables the assessment of schema matching tools [25]. It includes a classification of task-oriented datasets and new metrics for computing the post-match effort. STBenchmark aims at evaluating the quality of the mapping functions and their execution time [26]. Datasets are gathered according to common transformations (e.g., copying, flattening) but they can be enriched using instance generators, which can be tuned with configuration parameters (e.g., kinds of join, nesting levels).

The **entity matching** task, which is directly related to spatial entity matching, consists of discovering correspondences between equivalent objects. It also benefits from two benchmarks. The former proposes a set of four datasets about e-commerce and scientific publications [27]. These static datasets were used to compare entity matching tools. On the other hand, EMBench is based on importing existing entities (e.g., from Linked Open Data) and applying modifiers to their features (e.g., abbreviation, synonyms) [28]. These changes generate a set of modified entities, which form an entity matching dataset when grouped with the original entities. Although these two entity matching benchmarks are useful in most contexts, they are insufficient when dealing with spatial matching.

To the best of our knowledge, there is no benchmark for evaluating spatial entity matching approaches. Kang et al. propose a tool to detect the corresponding spatial entities [29]. They take two sources of entities as input, the potentially corresponding entity pairs being automatically detected based on a specified similarity measure. Then, the user has to make a decision for each pair to be considered as corresponding or not. This tool may be interesting to create a training dataset but it is not enough for a benchmark. Beeri et al. implemented a random-dataset generator to evaluate their spatial entity matching approach [30]. They generate two datasets of spatial entities in which some entities are corresponding. Unfortunately, generated entities are described only by spatial information (longitude and latitude) because their proposed matching approach exploits only the spatial information to detect the correspondences. Otherwise, the datasets used in the spatial entity matching papers are not made fully available, for instance because of confidentiality issues. A few attempts are available, such as a dataset about restaurants [31]. Yet, they cannot be exploited for various reasons. Some of them are not challenging (e.g., a simple equality metric applied to the phone numbers in the restaurant dataset discovers all the correct corresponding entities). In addition, a specific dataset may be required, for instance to include all POI types (e.g., restaurants, museums, mountains) or all entities from a given area. This lack of benchmark does not facilitate a fair and accurate comparison of the different spatial matching approaches. We also argue that the properties of a dataset are useful, both for understanding why a spatial entity matching approach is (not) effective, and for using appropriate training data when needed.

III. Taxonomy

In this section, we present a taxonomy to characterize the differences between the LBS providers. We start by introducing preliminary definitions that describe a model of LBS providers.

A. Preliminary definitions

It is necessary to understand the context of the LBS in order to construct a process to integrate them. In this section we illustrate a model that describes the LBS context of multi-provider.

Definition 1: Point of Interest (POI)
A POI is a geographical object described by a set of properties. Among these properties, there is a name, a type (e.g., restaurant, castle), a location (positioning coordinates) and a geometric shape (e.g., point, line, polygon). It is defined by the tuple:

$$POI = (name, type, coordinates, shape)$$

For example, the tuple ET below represents the Eiffel Tower POI:

$ET =$ (Eiffel Tower, Tourist, (48.858439, 2.294474), Point)

We distinguish between *large POI* and *small POI* relative to the POI area in the real-world. For example, a pub is a *small POI* while a park is a *large POI*. Let us consider the set $\mathbb{P} = \{p_1, ..., p_q\}$ that contains all the POIs of the real world where q is the number of POIs. Each LBS provider offers a set of entities that refer to a subset of existing POIs. Currently, the entities are represented with a point geometrical shape. Regarding the entities that refer to POIs with large areas, they are approximated by points such as computing their center of gravity [30]. The entities offered by a provider are derived from a specific schema of that provider.

Definition 2: Schema of provider
The schema $\mathbb{S}_k$ describes the structure of entities offered by the provider k. It is defined by:

$$\mathbb{S}_k = \mathbb{I}_k \cup \mathbb{L}_k \cup \mathbb{A}_k \cup \mathbb{B}_k$$

where

- $\mathbb{I}_k = \{id_k\}$ is an internal identifier attribute that represents a given entity for the provider k.
- $\mathbb{L}_k = \{\text{LONGITUDE}_k.label, \text{LATITUDE}_k.label\}$ is a pair of spatial attributes standing for the spatial information.
- $\mathbb{A}_k = \{\text{NAME}_k.label, \text{POITYPE}_k.label\}$ is a pair of terminological attributes that are mandatory. We call them primary attributes because they exist in the schemas of all providers and always have values.
- $\mathbb{B}_k = \left\{att\mathbb{B}_k^1.label, ..., att\mathbb{B}_k^r.label\right\}$ is another set of terminological attributes that are optionally provided with $r = |\mathbb{B}_k|$. We call them secondary attributes because they may be either missed from some schemas or have null values.

Hypothetically, a schema of any provider k includes at least all attributes in $\mathbb{I}_k \cup \mathbb{L}_k \cup \mathbb{A}_k$. We note att_k^i any attribute of the schema $\mathbb{S}_k$. The abstract data type of att_k^i, denoted as $att_k^i.type$, may have one of the following data types: string, number, array or associative array. Note that a schema may be static or dynamic. A static schema has fixed labels and structures, while for a dynamic schema, labels and structures can be modified. As an example, the provider *OpenStreetMap* [32] has a dynamic schema in which the user can add new attributes for some entities. In contrast, *GoogleMap* [33] has a static schema, so that the number and the labels of the attributes are common for all the entities. The entity set of a provider k is denoted by $\mathbb{E}_k = \{e_1, ..., e_n\}$ where n is the number of entities.

Definition 3: Entity of POI
An entity of a POI of a provider k, denoted by $e \in \mathbb{E}_k$, is an instance of the schema $\mathbb{S}_k$ and refers to one real-world POI $p \in \mathbb{P}$.

$$\begin{aligned} e = &\{(id_k.label, id_k.val), (\text{LATITUDE}_k.label, \text{LATITUDE}_k.val), \\ &(\text{LONGITUDE}_k.label, \text{LONGITUDE}_k.val), \\ &(\text{NAME}_k.label, \text{NAME}_k.val), (\text{TYPE}_k.label, \text{TYPE}_k.val), \\ &\left(att\mathbb{B}_k^1.label, att\mathbb{B}_k^1.val\right), ..., (att\mathbb{B}_k^r.label, att\mathbb{B}_k^r.val)\} \end{aligned}$$

where r is the number of secondary attributes of the schema $\mathbb{S}_k$. Table I shows an example of two entities x and y offered by two different providers that represent the POI ET (*Eiffel Tower*) with two different schemas. We denote $\mathbb{E} = \bigcup_{k=1}^{m} \mathbb{E}_k$, the union set of m providers' entities sets.

Definition 4: Association function f
The association function f is defined by:

$$\begin{aligned} f \colon &\mathbb{E} \to \mathbb{P} \\ &e \to f(e) = p \end{aligned}$$

such that $e \in \mathbb{E}$ refers to p.

For example, the entity x of Table I refers to the POI ET (*Eiffel Tower*) and $f(x) = ET$.

Definition 5: Corresponding entities
Two entities $e_1 \in \mathbb{E}_1$ and $e_2 \in \mathbb{E}_2$ are corresponding entities, denoted $e_1 \equiv e_2$, iff

$$\exists\, p \in \mathbb{P} \backslash f(e_1) = f(e_2) = p$$

For example, the two entities x and y of Table I are corresponding entities ($x \equiv y$) because they refer to the same POI ET (*Eiffel Tower*).

Definition 6: Corresponding attributes
Two attributes $att_1^i \in \mathbb{S}_1$ and $att_2^j \in \mathbb{S}_2$ are two corresponding attributes, denoted $att_1^i \equiv att_2^j$, iff they represent the same concept.

In the literature of schema matching, the correspondences between attributes are represented by a relationship [24] such as equivalence, overlap, disjointness, exclusion. But in the context of LBS providers, we only consider the equivalence relationship.

In the next section, we use the above definitions to introduce the taxonomy of differences.

B. *Taxonomy of differences*

In this section, we propose a formalization of the various differences that may arise between the entities of two providers. To illustrate the comparison between providers, let us consider $\mathbb{E}_1$ and $\mathbb{E}_2$ as entity sets of two LBS providers. Let $\mathbb{S}_1 = \mathbb{I}_1 \cup \mathbb{L}_1 \cup \mathbb{A}_1 \cup \mathbb{B}_1$ and $\mathbb{S}_2 = \mathbb{I}_2 \cup \mathbb{L}_2 \cup \mathbb{A}_2 \cup \mathbb{B}_2$ be the schemas of $\mathbb{E}_1$ and $\mathbb{E}_2$ respectively. Also, consider the POI $p \in \mathbb{P}$ and two corresponding entities $e_1 \in \mathbb{E}_1$ and $e_2 \in \mathbb{E}_2$ that refer to p (i.e., $e_1 \equiv e_2$).

The potentially corresponding entities of several sets will be compared depending on four levels: 1) schema, 2) terminology, 3) spatial and 4) entities' availability.

1) Schema Differences:
This level explains the heterogeneity between distinct schemas where two differences are distinguished.
Generally, differences between schemas involve two providers, i.e., Inter-Difference. In the case of a provider with a dynamic schema, differences may be classified as Intra-Difference.

Attribute Heterogeneity:
The attribute heterogeneity consists of two corresponding attributes belonging to two distinct schemas and have different labels or different abstract data types. In Table I, the attribute

TABLE I. EXAMPLE OF TWO ENTITIES x AND y, OFFERED BY TWO DIFFERENT PROVIDERS, THAT REPRESENT THE POI ET (*Eiffel Tower*) WITH TWO DIFFERENT SCHEMAS

Model	**Entity x (offered by provider 1)**	**Entity y (offered by provider 2)**
$\mathbb{I}$	EntityID : *51190385*	id : *fd0cfb424bbd79bf28a832e1764f1c2aa5927714*
$\mathbb{L}$	Latitude : *48,858606*	geometry : { location : { lat : *48.85837*,
	Longitude : *2,293971*	lng : *2.294481*}}
$\mathbb{A}$	DisplayName : *Tour Eiffel*	name : *Eiffel Tower*
	EntityTypeID : *7999*	types : *establishment*
$\mathbb{B}$	Phone : *0892701239*	formatted_phone_number : *+33 892 70 12 39*
	CountryRegion : *FRA*	website : *http://www.tour-eiffel.fr*
	Locality : *Paris*	formatted_address : *Champ de Mars,*
	PostalCode : *75007*	*5 Avenue Anatole France, 75007 Paris, France*
	AddressLine : *Champ De Mars, Avenue Anatole France ...*	...

DisplayName in the schema of the provider 1 and the attribute *name* in the schema of the provider 2 represent the name of the POI but they have different labels. Consider two attributes $att_1^i \in \mathbb{S}_1$ and $att_2^j \in \mathbb{S}_2$, $\mathbb{S}_1$ and $\mathbb{S}_2$ have an attribute heterogeneity difference iff

$$\left(att_1^i \equiv att_2^j\right) \wedge$$
$$\left(att_1^i.label \neq att_2^j.label \vee att_1^i.type \neq att_2^j.type\right)$$

Different Structures:
Schemas may have various structures. One attribute of one schema may correspond to two or more attributes of another schema. Returning to Table I, the address is represented by three attributes *Locality, PostalCode* and *AddressLine* in the schema of the provider 1 while the attribute *formatted_address* in the schema of the provider 2 represents the full address. That is, a concept is described by one attribute of the schema $\mathbb{S}_1$ and by two or more attributes of the schema $\mathbb{S}_2$, or vice versa.

$$att_1^i \equiv (att_2^1, att_2^2, \ldots) \vee (att_1^1, att_1^2, ...) \equiv att_2^j$$

There are complex correspondences between the structures of the schemas. For instance, more than one attribute of a schema may correspond to more than one attribute of another (i.e., [n:m] correspondences). We do not consider the complex correspondences in this paper since the schemas in the context of LBS are simple.

2) Terminological Differences:
This level is related to the heterogeneity of values for primary and secondary terminological attributes of two corresponding entities.

Different Data:
Two corresponding entities have different values for their corresponding terminological attributes (primary or secondary). e_1 and e_2 have different data iff

$$\exists\, att_1^x \in \mathbb{A}_1 \cup \mathbb{B}_1, \exists\, att_2^y \in \mathbb{A}_2 \cup \mathbb{B}_2 \backslash$$
$$e_1 \equiv e_2 \ \wedge (e_1.att_1^x \equiv e_2.att_2^y) \wedge$$
$$(e_1.att_1^x.val \neq e_2.att_2^y.val)$$

Note that the degree of difference between the data varies. This variation may be classified as semantic (SEM) or syntactic (SYN). The former consists of two corresponding attributes that have different values but are based on the same concept (e.g., *eat-drink* and *restaurant* are two POI types that have the same meaning). The latter is about the syntax of corresponding attributes' values. They are a consequence of the different ways that a value can be written in real life, without any alteration of its meaning, or a result of human errors (i.e., misspellings, word permutations, aliases, different standards, acronyms, abbreviations and multilingualism). In Table I, the type of the entity x is *7999* while it is *establishment* for the entity y. *Different Data* (SEM and SYN) is classified as Inter-Difference.

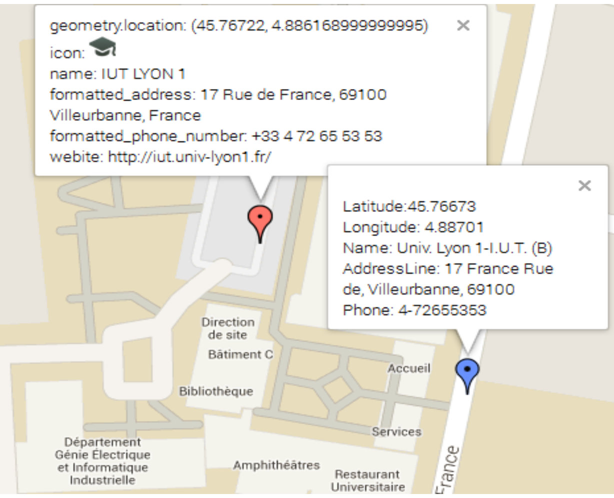
Missing Data (MD):
Two corresponding entities having a feature that is described by one entity and missed by the other. In Table I, the website is missing from the entity x while it is given by the entity y. This difference is classified as Inter-Difference. e_1 and e_2 have missing data iff

$$\Big(\exists\, att_1^x \in \mathbb{A}_1 \cup \mathbb{B}_1, \exists\, att_2^y \in \mathbb{A}_2 \cup \mathbb{B}_2 \backslash$$
$$(att_1^x \equiv att_2^y \ \wedge$$
$$(e_1.att_1^x.val = NULL \vee e_2.att_2^y.val = NULL))\Big)$$
$$\vee$$
$$\Big(\exists\, att_1^x \in \mathbb{A}_1 \cup \mathbb{B}_1, \forall\, att_2^y \in \mathbb{A}_2 \cup \mathbb{B}_2 \backslash$$
$$(att_1^x \not\equiv att_2^y)\Big)$$

Similar Data:
It consists of two entities that have similar values for terminological attributes but refer to two distinct POIs of the same type. Consider two POIs $p' \in \mathbb{P}$ and $p'' \in \mathbb{P}$ of the same type, and two entities $e' \in \mathbb{E}$ and $e'' \in \mathbb{E}$ that refer to p' and p'' respectively. If e' and e'' have similar values for any terminological attributes, then the difference between e' and e'' is denoted as similar data.

$$\exists\, att' \in \mathbb{B}', \exists\, att'' \in \mathbb{B}'' \backslash$$
$$(p' \neq p'') \wedge (p'.type = p''.type) \wedge (f\,(e') = p') \wedge$$
$$(f\,(e'') = p'') \wedge$$
$$((e'.\text{NAME}.val \cong e''.\text{NAME}.val) \vee$$
$$(e'.att'.val \cong e''.att''.val))$$

Usually, the *Similar Data* difference appears when we have two or more branches of the same organization. Entities that represent these branches are of the same type, have similar terminological values (e.g., place name), located in different



geometry.location: (45.76722, 4.886168999999995)
icon:
name: IUT LYON 1
formatted_address: 17 Rue de France, 69100 Villeurbanne, France
formatted_phone_number: +33 4 72 65 53 53
webite: http://iut.univ-lyon1.fr/

Latitude:45.76673
Longitude: 4.88701
Name: Univ. Lyon 1-I.U.T. (B)
AddressLine: 17 France Rue de, Villeurbanne, 69100
Phone: 4-72655353

Figure 1. Example of the *Equipollent Positions* difference: Two corresponding entities refer to the IUT Lyon 1 University (large POI) having different locations but both are correct.

areas and not corresponding to each other because each branch is a distinct POI. This differences is classified as Similar Data is classified as Intra-Difference and Inter-Difference.

3) Spatial Differences:
At this level, we investigate the problem of positioning between the corresponding entities. Three differences can be distinguished.

Different Locations (DL):
Two corresponding entities have different values for their corresponding spatial attributes. The two entities of Table I refer to *Eiffel Tower*, but they have different longitude and latitude values. The distance between the two locations is approximately 226 meters. This difference is classified as Inter-Difference.

$$e_1 \equiv e_2 \wedge (e_1.\text{LATITUDE}.val \neq e_2.\text{LATITUDE}.val \vee$$
$$e_1.\text{LONGITUDE}.val \neq e_2.\text{LONGITUDE}.val)$$

Equipollent Positions (EP):
This difference appears when the corresponding entities refer to a *large POI* and have different locations, but those locations are both correct with regard to the location of the POI that they represent. That is, corresponding entities' positions are equivalent in terms of concept but not in terms of values. This difference is classified as Inter-Difference.

$$\begin{aligned}
&(e_1 \equiv e_2) \wedge \\
&(e_1.\text{LONGITUDE}, e_1.\text{LATITUDE}) \subset p.coordinates \wedge \\
&(e_2.\text{LONGITUDE}, e_2.\text{LATITUDE}) \subset p.coordinates \wedge \\
&(e_1.\text{LONGITUDE}.val \neq e_2.\text{LONGITUDE}.val \vee \\
&e_1.\text{LATITUDE}.val \neq e_2.\text{LATITUDE}.val)
\end{aligned}$$

Figure 1 shows two corresponding entities refer to the *IUT Lyon 1 University* (large POI) and have different locations (center of gravity vs entrance gate) but both are correctly represented.

Superposition:
This consists of two entities that have the same locations but refer to two distinct POIs of the same type and it is classified as Intra-Difference and Inter-Difference. Usually, this case appears in shopping centers where two POIs of the same type are located one above the other on two different floors. Consider two POIs $p' \in \mathbb{P}$ and $p'' \in \mathbb{P}$ of the same type, and two entities $e' \in \mathbb{E}$ and $e'' \in \mathbb{E}$ that refer to p' and p'' respectively. If e' and e'' have the same location, then the difference between e' and e'' is denoted as superposition.

$$\begin{aligned}
&(p' \neq p'') \wedge (p'.type = p''.type) \wedge \\
&(f(e'') = p'') \wedge (f(e') = p') \wedge \\
&(e'.\text{LATITUDE}.val = e''.\text{LATITUDE}.val) \wedge \\
&(e'.\text{LONGITUDE}.val = e''.\text{LONGITUDE}.val)
\end{aligned}$$

4) Entity's Availability:
The entity's availability category takes into account errors that can be found in the entity set of a provider. Two differences can be distinguished at this level.

Not Found Entity:
This case, classified as Inter-Difference, consists of a POI that is given by one provider but not by the other. Considering the POI $p' \in \mathbb{P}$, p' is a Not found POI iff

$$\exists\, e_1 \in \mathbb{E}_1, \forall\, e_2 \in \mathbb{E}_2 \backslash f(e_1) = p' \wedge f(e_2) \neq p'$$

Duplicate Entities:
This case, classified as Intra-Difference, corresponds to two entities of the same provider that refer to the same POI. Consider two entities $e_1 \in \mathbb{E}_1$ and $e_1^{'} \in \mathbb{E}_1$, e_1 and $e_1^{'}$ are two duplicate entities iff

$$\exists\, p \in \mathbb{P} \backslash \left(f(e_1) = f\left(e_1^{'}\right) = p \right)$$

Although the differences described in the taxonomy are elementary, the detection of the corresponding entities requires some hard work because a combination of differences may occur when comparing two entities. For instance, the two entities x and y of Table I are two corresponding entities with a combination of four differences, namely 1) *Attribute Heterogeneity*, 2) *Different Structure*, 3) *Syntactic Different Data* (SYN) and 4) *Different Locations* (DL).

IV. Benchmark

The taxonomy of differences is useful to get statistics about LBS providers and to understand how they can be integrated. Also, it serves to create a benchmark that evaluates the performance of spatial entity matching approaches and to build a characterized dataset that serves for machine learning purposes. In this section we describe PABench [20] based on this taxonomy.

A. *Overview of the matching process*

In general, the data integration process consists of three consecutive phases namely 1) schema/ontology matching, 2) entity matching and 3) entity merging or fusion. The schema matching task helps finding corresponding attributes between two schemas in order to compare their values later in the entity matching task. It produces an alignment between the corresponding attributes accompanied by a transformation function such as combination, split, etc. A schema matching approach must be able to handle the differences denoted in the schema category namely *Attribute Heterogeneity* and *Different structures*. In the context of LBS, the schema matching of providers can be done manually since their schemas are small and simple, so that there is no need for semi- or fully-automatic approaches to handle the schema matching task. Secondly, (spatial) entity matching approaches are used to find corresponding entities in several datasets to merge them together. It takes as input the datasets that need to be merged and the alignment of their schemas produced by the schema matching task. Entity matching can be done by computing a similarity score between each pair of entities. Then, a matching approach considers a pair of entities as corresponding if its similarity score is higher than a given threshold [34], or produces a list of pairs of entities ranked according to their similarity score. Concerning spatial entity matching systems, they measure the degree of similarity between entities using various techniques such as Euclidean distance between entities' locations, semantic equivalence and syntactic comparison of terminological information. These measures serve to compute a score that indicates their belief that two entities correspond. Finally, the entity merging phase takes as input the result of the entity matching task in order to fuse the corresponding entities. How the entities are merged depends on how these entities will be used, it may be done by a simple combination of values or based on specific rules in a given context. In our context, we intend to consider the tourist rules to merge the spatial entities. Note that the schema matching and the entity merging are no longer discussed in this paper; next we focus on the evaluation of spatial entity matching approaches.

To evaluate the performance and the results' quality of a spatial entity matching system, consider two datasets, namely source and target, for which a list of correct correspondences, called ground truth, is known in advance. For each entity in the source dataset, the matching system will try to find the corresponding entity from the target dataset. Thus, correspondences returned by the matching system are compared to the ground truth correspondences in order to measure how successfully the matching system detects the expected answer.

B. Benchmark construction

PABench has been constructed based on the differences defined in the taxonomy. Recall that a spatial entity consists of spatial and terminological (primary and secondary) information and refers to a real world POI. Deciding whether two spatial entities correspond is a challenging task due to the differences that occur between them. As previously mentioned, two corresponding entities being compared may have a combination of differences where each combination is a distinct situation of differences. To understand the weaknesses and strong points of an entity spatial matching system, the evaluation must be characterized according to the situations of differences that may occur between entities. In other words, it is required to evaluate a spatial entity matching system based on each situation of differences.

The possible situations of differences are computed based on the taxonomy of differences with respect to the entity matching task. Since the entity matching goal is to detect the corresponding entities, only the differences concerning corresponding entities are considered, namely *Different Locations* (DL) and *Equipollent Positions* (EP) from the spatial category and *Missing Data* (MD), *Semantic Different Data* (SEM) and *Syntactic Different Data* (SYN) from the terminological category. *Superposition*, *Similar Data* and *Not Found Entity* differences may be used to add noise entities (see below) when comparing the source and target datasets. Finally, *Duplicate Entities* must be pre-handled before the entity matching task using deduplication techniques to ensure the quality of used datasets.

Spatial information is only expressed by an entity's location, it may have zero (i.e., no difference) or only one difference in the spatial category differences. The set of spatial differences S_dif is given by:

$$S\_dif = \{\emptyset, DL, EP\}$$

Primary terminological information is expressed by an entity's name and type, it may have zero, one (i.e., at least one attribute has the difference) or two differences of the terminological category differences. The *Missing Data* (MD) difference cannot be considered because the primary terminological attributes are always provided and have values (see Section III-A). The set of primary terminological differences PT_dif is given by:

$$PT\_dif = \{\emptyset, SEM, SYN, (SEM, SYN)\}$$

Secondary terminological information varies from one provider to another, it may have zero, one (i.e., at least one attribute has the difference), two (i.e., each difference appears at least once) or three differences of the terminological category differences. The set of secondary terminological differences ST_dif is given by:

$$\begin{aligned}ST\_dif = &\{\emptyset, SEM, SYN, MD, (SEM, SYN, MD),\\ &(SEM, SYN), (SEM, MD), (SYN, MD)\}\end{aligned}$$

Let $Situations\_dif$ be the set of all possible combinations of differences that may occur between two corresponding entities at all levels (spatial, primary terminological and secondary terminological)

$$\begin{aligned}Situations\_dif = &\{a, b, c \setminus a \in S\_dif, b \in PT\_dif,\\ &c \in ST\_dif\}\end{aligned}$$

where

$$|Situations\_dif| = 3 \times 4 \times 8 = 96$$

Returning to Table I, the two corresponding entities x and y have a combination of differences. For spatial information they have *Different Locations* (spatial coordinates), for primary terminological information they have *Syntactic Different Data* (name, type) and for secondary terminological information they have *Syntactic Different Data* (phone, address) and *Missing Data* (website). The situation $s \in Situations\_dif$ between x and y is given by $s = \{DL, SYN, (SYN, MD)\}$.

To guarantee that the situations have no redundancy, each situation $s \in Situations\_dif$ must be unique and exclusive, in the sense that the situations do not share any relation between them such as intersection, inclusion, etc. Thus, our benchmark consists of comparing a source dataset with a target dataset in which the corresponding entities have a specific situation of differences. If the correct answer

TABLE II. CONTINGENCY TABLE OF EVALUATION MEASURES.

Ground truth Matching approach	Corresponding entities	Non-corresponding entities
Corresponding entities	True Positive (TP)	False Positive (FP)
Non-corresponding entities	False Negative(FN)	True Negative (TN)

(represented by the ground truth) is returned by a matching approach, it means that it is able to deal with the given situation.

Definition 7: **TestCase**
For each situation $s \in Situations\_dif$, we define a test case that consists of a source dataset $\mathbb{E}_S \subset \mathbb{E}$ ($\mathbb{E} = \bigcup_{k=1}^{m} \mathbb{E}_k$), a target dataset $\mathbb{E}_T \subset \mathbb{E}$ and a ground truth between both source and target datasets.

$$TestCase(s) = (\mathbb{E}_S,\, \mathbb{E}_T,\, groundTruth)$$

Noise entities may be added to $\mathbb{E}_S$ and $\mathbb{E}_T$. A noise entity is an entity that exists in one dataset and does not have any correspondence in the other dataset. The goal of adding noise entities is to explain whether a matching approach is able to avoid detecting two non-corresponding entities even if they have near locations or similar information. Noise entities should contain entities with *Not Found Entity*, *Similar Data* and *Superposition* differences. Concerning the *Not found Entity* difference, it can be easily detected from the real entities of LBS providers. But *Similar Data* and *Superposition* are hard to detect, in this case we intend to automatically generate entities with such differences. These test cases allow us to find the situations of differences that a matching approach is able to handle and to what degree this handling is possible in order to differentiate it from other similar approaches.

C. Quality measure and impact of differences

Results' quality of a matching system is measured by the standard performance measures that come from the information retrieval domain, precision, recall and F-measure [35]. These measures evaluate the performance of a matching system by comparing its results to ground truth's results. Also, they help to understand weaknesses and strengths of a matching approach for each test case. Table II classifies the contingency of evaluation measures' base. Precision calculates the proportion of correct correspondences detected by the matching system among all detected correspondences. Using the notations of Table II, the precision is given by formula (1). A 100% precision means that all correspondences detected by the matching system are true.

$$precision = \frac{TP}{TP + FP} \qquad (1)$$

Recall computes the proportion of correct correspondences detected by the matching system among all correct correspondences. The recall is given by formula (2). A 100% recall means that all correct correspondences have been found by the matching system.

$$recall = \frac{TP}{TP + FN} \qquad (2)$$

F-measure is a trade off between precision and recall and it is calculated with the formula (3). The β parameter of formula (3) regulates the respective influence of precision and recall ($\beta \in \mathbb{R}^+$). It is often set to 1 to give the same weight to these two evaluation measures.

$$F - measure\,(\beta) = \frac{(\beta^2 + 1) \times precision \times recall}{(\beta^2 \times precision) + recall} \qquad (3)$$

It is important to analyze how the differences of the taxonomy impact these measures in order to discover the weaknesses and the strengths of an approach. Concerning the *Attribute Heterogeneity*, *Different Structure* and *Duplicate Entities* differences, they are pre-handled before launching the spatial entity matching process. *Different Data* (SEM and SYN), *Missing Data*, *Different Locations* and *Equipollent Positions* must be addressed through the matching system. They impose obstacles that may prevent a matching system from detecting the correct corresponding entities. Hence, if a matching approach fails to overcome those obstacles, True Positive (TP) decreases, False Negative (FN) increases and recall decreases. Concerning *Superposition* and *Similar Data*, two distinct entities with the same location or with similar data may be detected as corresponding. These differences increase False Positive (FP) and precision decreases. The *Not Found Entity* concerns a POI that is represented by one provider and not by the other, that means the entity of the first provider does not correspond to any of the entities of the second provider. But it risks a situation where a matching approach detects a correspondence for an entity of this difference, which increases FP and precision decreases. Also, this difference impacts the entity merging phase because as long as the number of available entities is small, we cannot ensure the correctness of information. In the case of *Duplicate Entities*, a matching approach may detect the same correspondence twice. This case will increase TP leading to a wrong precision value. That is why it is important to verify the quality of providers' datasets before starting the matching process using deduplication techniques. Table III summarizes the taxonomy of differences and their impacts on the quality measures.

V. DATASETS

A tool that consists of two modules has been implemented in order to generate the datasets of test cases. The first module, called GeoBench [36], is addressed to experts and it serves to build a characterized dataset in a semi-automatic process through the sets of three LBS providers namely Google Maps, Nokia Here Maps and Geonames. Let $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$ be the datasets of the three LBS providers respectively. Experts can search for a specific or random source entity from one provider, and then GeoBench searches for the nearby target entities from the two others separately. For each retrieved target entity, an expert has to decide whether it corresponds to the source entity and to select the differences that exist between the two entities at each level (spatial, primary and secondary terminological). Concerning the secondary terminological level, only the most common secondary information is considered namely phone number, website and address. These three secondary attributes may have *Missed Data* (MD) or *Syntactic Difference Data* (SYN) assuming that it is impossible to have a *Semantic Different Data* (SEM) according to the information that they represent (e.g., two corresponding phone numbers may be

TABLE III. TAXONOMY OF DIFFERENCES AND QUALITY MEASURES IMPACT.

Category	Difference	Intra-Diff	Inter-Diff	Impact
Schema	Attribute Heterogeneity	X(dynamic schema)	X	
	Different structure	X(dynamic schema)	X	
Terminology	Different Data (SEM)		X	TP ↘ FN ↗
	Different Data (SYN)		X	TP ↘ FN ↗
	Missing Data (MD)		X	TP ↘ FN ↗
	Similar Data (SD)	X	X	FP ↗
Spatial	Different locations (DL)		X	TP ↘ FN ↗
	Equipollent Positions (EP)		X	TP ↘ FN ↗
	Superposition (SUP)	X	X	FP ↗
Availability	Not found POI		X	FP ↗
	Duplicate Entities	X		TP (wrong value)

TABLE IV. NUMBER OF ENTITIES AND CORRESPONDENCES PRODUCED USING GEOBENCH (OCTOBER 2014).

Dataset	Number of entities
$\mathbb{E}_1$	715
$\mathbb{E}_2$	583
$\mathbb{E}_3$	282
Total	**1580**

Number of correspondences	
$\mathbb{E}_1$, $\mathbb{E}_2$	569
$\mathbb{E}_1$, $\mathbb{E}_3$	254
$\mathbb{E}_2$, $\mathbb{E}_3$	247
Total	**1070**

TABLE V. TOP FIVE TEST CASES ACCORDING TO THE NUMBER OF DETECTED CORRESPONDENCES (OCTOBER 2014).

Test Case #	Situation	Number of Correspondences
43	{EP, SYN, {SYN, MD}}	145
27	{DL, SYN, {SYN, MD}}	78
41	{EP, SYN, SYN}	69
35	{EP, SEM, {SYN, MD}}	62
11	{0, SYN, {SYN, MD}}	60

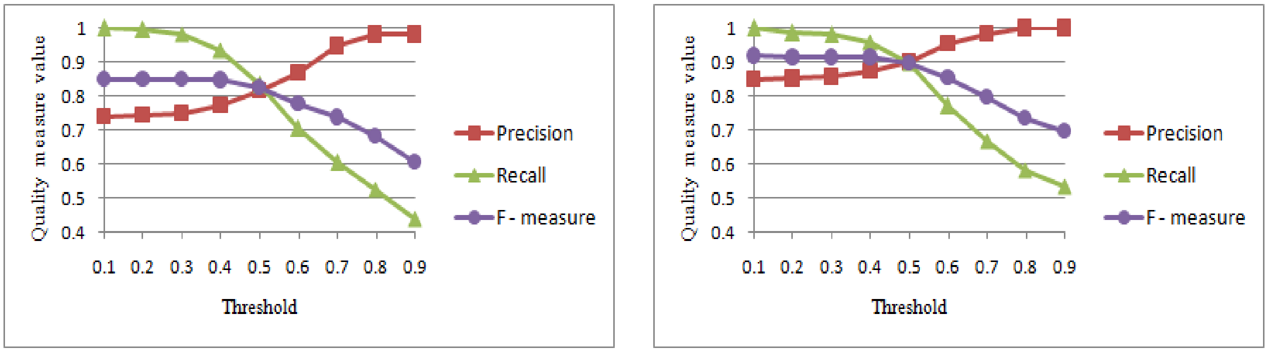
syntactically different but can never be semantically different). In this case, the number of the possible situations decreases from 96 to 48. In other contexts, the *Semantic different Data* (SEM) may be considered (e.g., when comparing the food type of two entities of restaurant's type, Pizza vs Italian food). GeoBench allows us to create a dataset $\mathbb{E} = \mathbb{E}_1 \cup \mathbb{E}_2 \cup \mathbb{E}_3$ in which, for each pair of entities, we know the relevance of correspondence and the situation of differences. Note that $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$ contain only the entities processed by GeoBench and not the whole set of entities of the three LBS providers.

The second module has been implemented to generate the test cases, it uses the dataset $\mathbb{E}$ created by GeoBench to generate source and target datasets for each situation $s \in Situations\_dif$. This module allows to configure the characteristics of a test case through a set of parameters in order to control aspects such as situation of differences, percentage of correspondences and percentage of noise. Once the characteristics of a test case are configured, source and target datasets are generated with a ground truth file so the evaluators can assess the results of their approaches. The module will search for all pairs of entities that match the requested situation of differences, then the entities of each pair will be distributed between the source and the target datasets and, the identifiers of corresponding entities will be listed together in the ground truth file. Finally, noise entities will be added to the target dataset or source dataset.

Statistics shown in Table IV represent the number of entities of each provider's dataset and the number of correspondences between them. Table V provides the top five test cases according to the number of detected correspondences. More statistics are available online along with PABench[20]. Retrieved datasets describe real world POIs where entities have been retrieved from several existing LBS providers using their web services. These datasets do not contain any redundancies or duplicated entities. The current version of the benchmark, as of October 2014, contains 1070 corresponding entities. All entities are available in CSV/SQL standard format to be easily parsed and used. To ensure that the test cases are rich enough, entities are distributed in several geographical zones/countries and refer to POIs of several types including *large POIs* and *small POIs*. Note that in the practice of the LBS context, some situations of differences rarely occur (e.g., the situation where two entities from different providers have no difference at all). In the future, we intend to develop an entity generator tool that takes a subset of source entities to modify the values of their attributes (spatial, primary and secondary terminological) in order to create a target dataset that expresses these rare situations.

VI. ANALYZE AND FIRST USE OF PROVIDERS' DATASETS

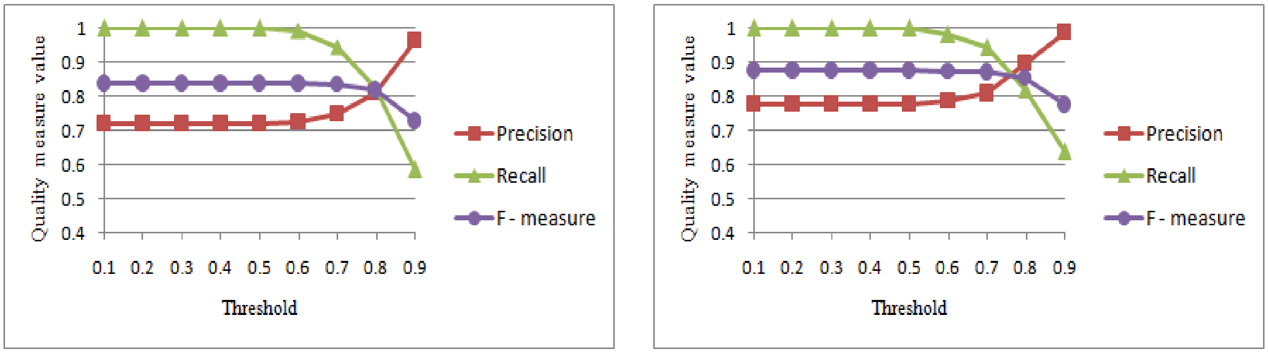
In this section, we demonstrate the resistance of the benchmark against frequently used basic measures. To reach this goal, a simple matching tool (one similarity measure and a threshold) is used to determine the difficulty of the spatial entity matching task and to show the heterogeneity of our collected dataset. Our basic matching tool consists in comparing the values of a single terminological attribute using a string similarity measure. Entity pairs that have the highest similarity score above a given threshold are considered as corresponding. This simple approach is used to match $\mathbb{E}_1$ with $\mathbb{E}_2$ and $\mathbb{E}_1$ with $\mathbb{E}_3$. Concerning the terminological information, we will compare the values of the NAME attribute using Levenshtein string similarity measure [37]. Experiments are repeated by varying the threshold value. The results are measured in terms of precision, recall and F-measure (see Section IV-C). Figures 2(a) and 2(b) show the matching quality of $\mathbb{E}_1$ with $\mathbb{E}_2$ and $\mathbb{E}_1$ with $\mathbb{E}_3$ respectively. The x-axis represents the values of threshold and the y-axis represents the values of precision, recall and F-measure. For a small value of threshold (0.1), the precision is low (75% for $\mathbb{E}_1$ vs $\mathbb{E}_2$ and 85% for $\mathbb{E}_1$ vs $\mathbb{E}_3$), which means that 25%-15% of detected corresponding entities are not correct according to the ground truth. However, the recall is high (99% in both cases), which means that the matching approach does not miss any of the ground truth correspondences. Increasing the threshold increases the precision and decreases the recall. For a high value of threshold (0.9), precision increases up to 98% in both cases, which indicates that most of the detected



(a) Dataset E1 - E2

(b) Dataset E1 - E3

Figure 2. Results' quality in terms of precision, recall and F-measure using Levenshtein similarity measure.



(a) Dataset E1 - E2

(b) Dataset E1 - E3

Figure 3. Results' quality in terms of precision, recall and F-measure using JaroWinkler similarity measure.

TABLE VI. EXECUTION TIME FOR LEVENSHTEIN AND JAROWINKLER ACCORDING TO THE NUMBER OF ENTITIES IN MATCHED DATASETS.

Datasets	Number of comparisons	Execution time of Levenshtein (sec)	Execution time of JaroWinkler (sec)
$\mathbb{E}_1$ vs $\mathbb{E}_2$	715 × 583	11	840
$\mathbb{E}_1$ vs $\mathbb{E}_3$	715 × 282	5	400
$\mathbb{E}_2$ vs $\mathbb{E}_3$	583 × 282	4	320

correspondences are correct according to the ground truth. Conversely, the recall decreases to approximately 50%, which means that the half of the ground truth correspondences are missing. A trade-off is reached for a 0.5 threshold, in which case the *F-measure* is up to 80% for $\mathbb{E}_1$ vs $\mathbb{E}_2$ and 90% for $\mathbb{E}_1$ vs $\mathbb{E}_3$, which indicates that using a simple matching method is not enough to resolve the matching problem. Similar results are obtained when the Levenshtein measure is replaced by the JaroWinkler measure [38][39][40] (see Figures 3(a) and 3(b)). A trade off is achieved with a 0.8 threshold and the *F-measure* is up to 80% for $\mathbb{E}_1$ vs $\mathbb{E}_2$ and 87% for $\mathbb{E}_1$ vs $\mathbb{E}_3$. Except that JaroWinkler keeps the same quality with threshold up to 0.6 while it decreases with Levenshtein from 0.4, this means that the scores calculated by Levenshtein for corresponding entities is lower than the scores calculated by JaroWinkler. The two similarity measures are approximately equivalent in terms of results' quality, but the Levenshtein metric is more efficient in terms of execution time. Table VI compares their execution times according to the number of entities in matched datasets.

These experiments show that basic similarity measures are not enough to match the real and heterogeneous data of PABench. Note that matching $\mathbb{E}_2$ with $\mathbb{E}_3$ have the same trend as matching $\mathbb{E}_1$ with $\mathbb{E}_2$ and $\mathbb{E}_1$ with $\mathbb{E}_3$.

VII. CONCLUSION

Spatial entity matching has become a basic problem in many application domains such as heterogeneous location-based services. In this paper, we highlighted the absence of a benchmark to compare and evaluate spatial entity matching approaches. We proposed a taxonomy that characterizes differences, heterogeneities and errors between LBS providers at four levels: schema, terminology, spatial and availability. We studied the impact of the identified differences on the results' quality of a matching approach and we proposed the necessary specifications to design a benchmark, called PABench, that serves to evaluate and compare spatial entity matching approaches. We believe that our proposition will allow researchers to better evaluate their matching approaches, identify the capabilities of their approaches, and also guide performance improvements in existing spatial entity matching approaches. In the future, PABench may be extended by 1) adding more entities and 2) automatically generating entities to cover the situations of differences that occur only rarely in the LBS context. Also, we intend to create a survey that compares and evaluates existing approaches in terms of results' quality and execution time using our benchmark. This evaluation will explain the weaknesses and the strengths of current works, which will help to propose a better matching approach. On the other hand, the proposed taxonomy is limited to punctual geographical objects, but it may be extended to cover complex objects (e.g., polygons and lines) in order to be used for complex geographical data.

Acknowledgment

This work was supported by the LABEX IMU (ANR-10-LABX-0088) of Université de Lyon, within the program ”Investissements d'Avenir” (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). Authors would like to thank Google Maps, Geonames, Open Street Map and Nokia Here Maps for providing POIs' entities.

References

[1] I. N. Gregory, “Time-variant gis databases of changing historical administrative boundaries: A european comparison,” Transactions in GIS, vol. 6, no. 2, 2002, pp. 161–178.

[2] M. A. Cobb, F. E. Petry, and K. B. Shaw, “Fuzzy spatial relationship refinements based on minimum bounding rectangle variations,” Fuzzy Sets and Systems, vol. 113, no. 1, 2000, pp. 111–120.

[3] M. L. Casado, “Some basic mathematical constraints for the geometric conflation problem,” in Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisboa, Instituto Geogrfico Portugus, 2006, pp. 264–274.

[4] J. J. Ruiz, F. J. Ariza, M. A. Ureña, and E. B. Blázquez, “Digital map conflation: a review of the process and a proposal for classification,” International Journal of Geographical Information Science, vol. 25, no. 9, 2011, pp. 1439–1466.

[5] A. Saalfeld, “A fast rubber-sheeting transformation using simplicial coordinates,” The American Cartographer, vol. 12, no. 2, 1985, pp. 169–173.

[6] C.-C. Chen, S. Thakkar, C. A. Knoblock, and C. Shahabi, “Automatically annotating and integrating spatial datasets,” in Advances in Spatial and Temporal Databases, 2003, pp. 469–488.

[7] S. Volz, “An iterative approach for matching multiple representations of street data,” in Proceedings of the JOINT ISPRS Workshop on Multiple Representations and Interoperability of Spatial Data, Hannover, 2006, pp. 101–110.

[8] A. Saalfeld, “Conflation automated map compilation,” International Journal of Geographical Information System, vol. 2, no. 3, 1988, pp. 217–228.

[9] Y. Doytsher, “A rubber sheeting algorithm for non-rectangular maps,” Computers & Geosciences, vol. 26, no. 9, 2000, pp. 1001–1010.

[10] M. Zhang, W. Shi, and L. Meng, “A generic matching algorithm for line networks of different resolutions,” in Workshop of ICA Commission on Generalization and Multiple Representation Computering Faculty of A Coruña University-Campus de Elviña, Spain, 2005.

[11] E. M. Arkin, L. P. Chew, D. P. Huttenlocher, K. Kedem, and J. S. B. Mitchell, “An efficiently computable metric for comparing polygonal shapes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 3, 1991, pp. 209–216.

[12] M. Gombosoi, B. Zalik, and S. Krivograd, “Comparing two sets of polygons,” International Journal of Geographical Information Science, vol. 17, no. 5, 2003, pp. 431–443.

[13] A. Masuyama, “Methods for detecting apparent differences between spatial tessellations at different time points,” International Journal of Geographical Information Science, vol. 20, no. 6, 2006, pp. 633–648.

[14] S. Thakkar, C. A. Knoblock, and J. L. Ambite, “Quality-driven geospatial data integration,” in ACM International Symposium on Geographic Information Systems, Washington, USA, 7-9 November, 2007, p. 16.

[15] E. Safra, Y. Kanza, Y. Sagiv, C. Beeri, and Y. Doytsher, “Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets,” International Journal of Geographical Information Science, vol. 24, no. 1, 2010, pp. 69–106.

[16] A.-M. O. Raimond and S. Mustire, “Data matching - a matter of belief,” in International Symposium on Spatial Data Handling (SDH), 2008, pp. 501–519.

[17] R. Karam, F. Favetta, R. Kilany, and R. Laurini, “Integration of similar location based services proposed by several providers,” in Networked Digital Technologies, 2010, pp. 136–144.

[18] V. Sehgal, L. Getoor, and P. Viechnicki, “Entity resolution in geospatial data integration.” in ACM International Symposium on Geographic Information Systems, 2006, pp. 83–90.

[19] B. Berjawi, E. Chesneau, F. Duchateau, F. Favetta, C. Cunty, M. Miquel, and R. Laurini, “Representing uncertainty in visual integration,” in Proceedings of the 20th International Conference on Distributed Multimedia Systems, Pittsburgh, USA, 27-29 August, 2014, pp. 365–372.

[20] “PABench,” URL: http://liris-unimap01.insa-lyon.fr/benchmark [accessed: 2014-12-03].

[21] J. Euzenat and P. Shvaiko, Ontology matching. Heidelberg, Germany: Springer-Verlag, 2007, ISBN: 3-540-49611-4.

[22] “Ontology Alignment Evaluation Initiative,” URL: http://oaei.ontologymatching.org [accessed: 2014-12-03].

[23] J. Euzenat, M.-E. Rosoiu, and C. Trojahn, “Ontology matching benchmarks: generation, stability, and discriminability,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 21, 2013.

[24] Z. Bellahsene, A. Bonifati, and E. Rahm, Schema Matching and Mapping. Heidelberg, Germany: Springer-Verlag, 2011, ISBN: 978-3-642-16517-7.

[25] F. Duchateau and Z. Bellahsene, “Designing a benchmark for the assessment of schema matching tools,” in Open Journal of Databases (OJDB), vol. 1, no. 1. RonPub, Germany, 2014, pp. 3–25.

[26] B. Alexe, W. C. Tan, and Y. Velegrakis, “Stbenchmark: towards a benchmark for mapping systems,” Proceedings of the VLDB, vol. 1, no. 1, 2008, pp. 230–244.

[27] H. Köpcke, A. Thor, and E. Rahm, “Evaluation of entity resolution approaches on real-world match problems,” PVLDB, vol. 3, no. 1, 2010, pp. 484–493.

[28] E. Ioannou, N. Rassadko, and Y. Velegrakis, “On generating benchmark data for entity matching,” Journal on Data Semantics, vol. 2, no. 1, 2013, pp. 37–56.

[29] H. Kang, V. Sehgal, and L. Getoor, “Geoddupe: A novel interface for interactive entity resolution in geospatial data,” in International Conference on Information Visualisation, 2007, pp. 489–496.

[30] C. Beeri, Y. Doytsher, Y. Kanza, E. Safra, and Y. Sagiv, “Finding corresponding objects when integrating several geo-spatial datasets,” in ACM International Workshop on Geographic Information Systems, 2005, pp. 87–96.

[31] “Restaurants' dataset,” URL: http://cs.utexas.edu/users/ml/riddle/data.html [accessed: 2014-12-03].

[32] “OpenStreetMap,” URL: http://www.openstreetmap.org [accessed: 2014-12-03].

[33] “Google Maps,” URL: http://maps.google.com [accessed: 2014-12-03].

[34] J. Madhavan, P. A. Bernstein, and E. Rahm, “Generic schema matching with cupid,” in VLDB, 2001, pp. 49–58.

[35] C. J. V. Rijsbergen, Information Retrieval, 2nd ed. Newton, Massachusetts, USA: Butterworth-Heinemann, 1979, ISBN: 0408709294.

[36] A. Morana, T. Morel, B. Berjawi, and F. Duchateau, “Geobench: a geospatial integration tool for building a spatial entity matching benchmark,” in ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, USA, 2014, in press.

[37] V. Levenshtein, “Binary codes capable of correcting deletions, insertions and reversals,” Soviet Physics Doklady, vol. 10, no. 8, 1966, pp. 707–710.

[38] M. A. Jaro, “Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida,” Journal of the American Statistical Association, vol. 84, no. 406, 1989, pp. 414–420.

[39] W. E. Winkler, “The state of record linkage and current research problems,” in Statistical Research Division, US Census Bureau. Citeseer, 1999. [Online]. Available: http://www.census.gov/srd/www/byname.html

[40] M. A. Jaro, “Probabilistic linkage of large public health data files,” Statistics in medicine, vol. 14, no. 5-7, 1995, pp. 491–498.

Development of a System Prototype for Insurance Rate Information Based on Natural Disaster Risk

Jihun Kang, Insu Lee, and Jung Ok Kim
Spatial Information Research Institute
Korea Cadastral Survey Corporation
Seoul, Korea
{kangdaejang, les05, jungok}@lx.or.kr

***Abstract*—In Korea, a research has been conducted on individual insurance rates for natural disasters. However, there is no system yet for managing comprehensive disaster insurance rate data or for representing an insurance rate map. In this study, a system for preparing Natural Disaster Insurance rate maps is proposed to transparently operate and activate the natural disaster insurance. Then, the availability of the insurance rate management is confirmed to develop a prototype system based on the aforementioned architecture.**

Keywords-Insurance rate; insurance rate system; natural disaster; disaster map; open source.

I. INTRODUCTION

Over the last decade, Korea's financial damage from natural disasters has reached 8.3293 trillion won. Such damage has been caused mainly by floods due to heavy rains (51.6%), typhoons (26.3%), heavy snowfall (20.2%), and wind/gale (1.9%) [1]. For the victims to effectively recover from property losses in all kinds of storm and flood damages, Korea developed a Natural Disaster Insurance and has been operating it since 2008 [2].

Natural Disaster Insurance, which is controlled by the National Emergency Management Agency and is sold and operated by private insurance companies, charges a single rate across all regions [3]. To investigate the viability of this single-rate issue, a study on the application of individual insurance rates by region is ongoing. Also, in 2014, a Natural Disaster Insurance Rate Map was produced based on a study in Ulsan, South Korea.

Since individual insurance rates are calculated considering the risks (caused by typhoons, flooding, and snowfall) of the concerned area all together, they can be complex and variable. Because the Natural Disaster Insurance Scheme should be fair and reliable, it must be managed using a system. In this study, to verify the need for such a system, a web-based system prototype for the Insurance Rate information management was developed. Moreover, the availability of the system was confirmed by overlaying various spatial data on the insurance rate map and suggesting a system construction plan.

II. INSURANCE RATE INFORMATION SYSTEM PROTOTYPE

A web-based prototype system for a Natural Disaster Insurance rate information service was developed. The architecture of the system design is shown in Figure 1. The design of the architecture based on open-source software can be advantageous in cost and ease of use.

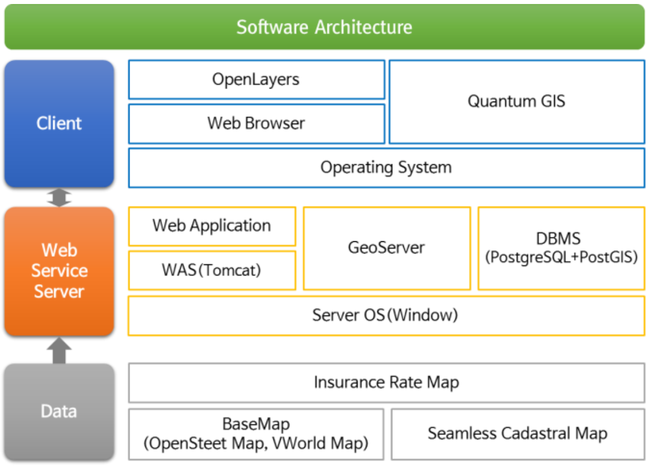


Figure 1. Software architecture of the prototype system

The Client tier consists of the web browser for displaying the information, Openlayers as the 2D Geographic Information System (GIS) engine, and Quantum GIS for managing the natural disaster risk data (insurance rate map) and the map-based data such as Digital Elevation Model (DEM) and the digital map. Although Quantum GIS is not directly related to the web-based prototype system, it is needed for the modification of the data such as for the projection transformation. The Server tier consists of Tomcat [5] for processing web applications, Geoserver for servicing geospatial data, and PostGIS (included in the PostgreSQL [6]) for managing the spatial and non-spatial databases. The Data tier consists of the insurance rate map based on natural disasters, the base map using the maps Application Programming Interface (API), and the seamless /cadastral map of Korea.

The insurance rate map is formed in the SHP data format, with the data coming from Ulsan, Korea. The base maps used were Open Street Map as the vector map, and VWorld Map as the aerial image map. VWorld Map is a map manufactured under the sponsorship of the Ministry of Land, Transport, and Maritime Affairs, which provides high-resolution aerial images in Korea. The seamless cadastral map is also formed in the SHP (ESRI shpapefile) data format and uses data from the same region as that for the insurance rate map. For the maps to be overlapped in between, ITRF 2000 Datum [7] and TM projection [8] was used.

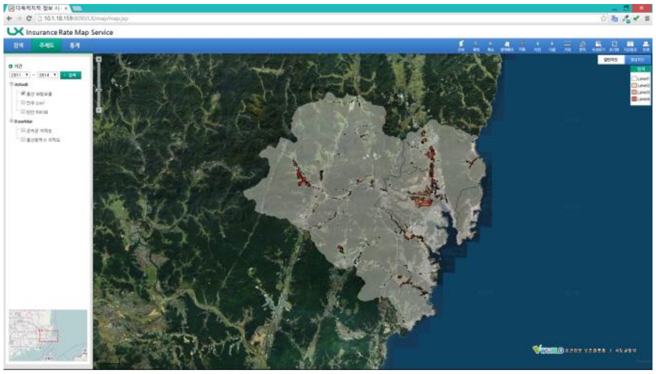


Figure 2. Main interface for the Insurance Rate Map service

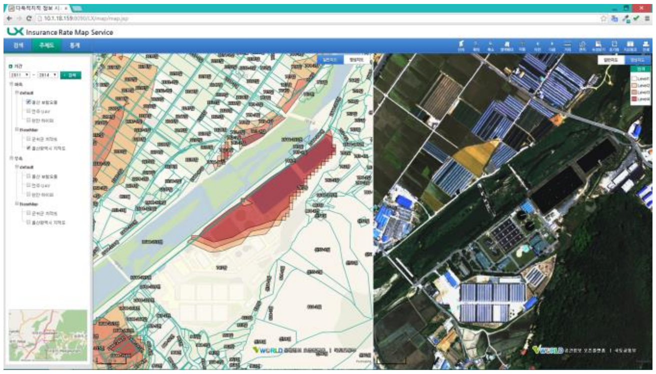


Figure 3. Geolink

The main interface of the prototype system is presented in Figure 2. This system can operate the general functions of map systems such as zoom, move, measurement tools, property view, and print. The system can also overlay a natural disaster map on a cadastral map for a nested view. In this case, the location of the nesting area is difficult to confirm, so Geo-link is implemented using Openlayers [9], as shown in Figure 3. Geo-link is a technology for splitting a browser window into two, with the locations of each half-window related to each other. Each layer of the respective domain can be individually controlled.

III. CONCLUSION AND FUTURE WORK

Through the web-based Natural Disaster Insurance information system prototype, the necessity of the insurance rate map service was assured. Also, the effective visualization of the insurance rate information was studied through overlay of spatial information. If the system is put in service, it will be able to operate a fair insurance scheme.

Although this prototype system is limited to providing premium rate information, in the near future through research, the National Insurance Rate Map is expected to be produced, which is based on Storm Risk Map, Flood Risk Map, and Snow Risk Map. To the extent of the production, a managing system to make them integrated is also needed. This management system will be used to predict and reduce natural disaster risks.

ACKNOWLEDGMENT

This research was supported by a grant for the ‘Development of an Insurance Rate Map Based on National Disaster Risk’ (NEMA-NH-2013-62) from the Natural Hazard Mitigation Research Group of the National Emergency Management Agency of Korea.

REFERENCES

[1] S. H. Nam, Korea Disaster Annual Report 2013, National Emergency Management Agency, pp. 580-582, 2013.

[2] National Disaster Information Center (http://www.safekorea.go.kr/dmtd/main/SdiwMain.jsp?q_menuid=M_NST_SVC_08), accessed 10 October 2014.

[3] J. S. Lee, I. S. Lee, J. H. Kang, and S. H. Yeon, GIS Technique for Analysis of Storm and Flood Insurance Premium Rates, Korean Society of Hazard Mitigation, pp. 271-350, 2014

[4] FIRM (http://www.fema.gov/floodplain-management/ flood-insurance-rate-map-firm), accessed 10 October 2014.

[5] Tomcat (http://tomcat.apache.org/), accessed 10 October 2014.

[6] PostgreSQL (http://www.postgresql.org/), accessed 10 October 2014.

[7] C. Boucher, Z. Altamimi, P. Sillard, and M. Feissel-Vernier, IERS Technical Note, IERS ITRS Centre, No. 31, 2004

[8] Transverse Mercator projection (http://en.wikipedia.org/wiki/Transverse_Mercator_projection), accessed 10 October 2014.

[9] Openlayers (http://openlayers.org/), accessed 10 October 2014.

ERP Spatial Management for Oil and Electric Networks

Ayman Ahmed Sami
KOC GIS unit (Contractor)
Openware (ESRI Official Distributer)
Ahmadi, Kuwait
engayman79@yahoo.com

Ahmad M Al-Kandari
College of Technological Studies
PAAET
Kuwait, Kuwait
alkandari1@yahoo.com

Ahmed Sami Abd Al Salam
Egyptian Electric Utility Regulatory Agency
Egyptian Electrical Municipality
Cairo, Egypt
dasaygd12@hotmail.com

***Abstract*—VCDM (Virtual Common Data Model) concept is used to accomplish integration between heterogonous and various workflows for enterprise organizations Virtual Common Data. The common elements between the heterogonous workflows are identified, creating SDM (Spatial Data Model) to act as the common data model based on analyzing the activities and workgroups for common processes in various workflows. Also, this paper shows how to analyze and develop the spatial data model to act as the unified geo framework for the various and heterogonous workflows in the enterprise organizations.**

Keywords-spatial asset management; location ID; spatial risk score matrix geo-framework; emergency planning zones.

I. INTRODUCTION

Williams [6] shows that many utilities in the practice of managing information remains undisciplined and unfocused. This can be seen by several aspects. It shows for example, at high costs and locked in resources associated with the development and redundant maintenance. Also, it is seen at overlapping point-to-point interfaces, an unmanaged glut of information (such as an accelerating volume and velocity of information sources) and an inflexibility of system design. All of these aspects cause a delay in response to evolving business needs. Utility IT (Information Technology) personals, engineers and operations leaders have recognized the need for asset data management. They recognize the value of enterprise GIS (Geographic Information System) as the required vehicle for asset data management.

Sami [4] presents that the major benefit to be derived from using GIS is the integration of disparate information systems, many of which already exist but which cannot communicate or exchange data effectively.

EPR (Emergency Response Planning) systems have been deployed with an overarching purpose to reduce the costs by managing processes and materials. The principal benefit of a consolidated ERP strategy was to ensure that data was not duplicated across departments, eliminating “islands of information”. Once the processes are separated, then they can now be linked, enabling the enterprise for wide planning and optimization. The changed competitive landscape will demand that ERP supports all goals regardless of their locations in an organization’s value network. Only then, enterprises can mobilize quickly and respond effectively to events as they occur at breakneck speed.

One of the main strategic objectives in the enterprise organizations is to develop the media, which can be used to analyze and show the impact of various and heterogonous workflows for achieving the scope with minimum time and cost and most attainable quality. VCDM is the approach which we will follow. Creation of VCDM is very challenging when we want to see the impact of completely different management plans within the organization. The following different systems and plans (impact of engineering system, field’s development system, exploration system, human resources and training management plans) have effects on each other. Combining the data from these systems will enable the user to reach the strategic organizational objectives. This challenge could be resolved through the SDM concept.

SDM concept is based on management and analysis of heterogeneous workflows through the spatial integrity in GIS environment. SDM relies on spatial modeling and analyzing workflow work packages or activities for planning, monitoring and controlling process for various business workflows. Analysis and implementation of workgroups of common processes for various workflows are more scalable and accurate than relying on common elements. It is difficult to identify the common elements among the heterogonous workflows, whereas analyzing the activities of processes common among these workflows could overcome this issue.

The state of the art is to form the SDM model. This SDM model will overcome the complexity of the existing approaches. It reduces maintenance and operation costs. Also, the responses are fast .

The critical point in SDM approach is creation of the spatial correlation between the analyzed and decomposed activities. The most efficient approach is forming the spatial risk data model. The decomposed work groups from various workflows can be spatially analyzed and managed, and then correlated using spatial risk data model. This paper presents a method to show how GIS can be utilized and managed as a unified geo framework through the development of spatial data model for ERP (Emergency Response Plan) for Oil and Electric Networks. The first section in this paper is about the spatial management of ERP for H2S (Hydrogen Sulphide) dispersion. The second part is about the spatial model of electric power networks in risk assessment management.

II. GIS AS UNIFIED GEO FRAMEWORK FOR EPZ (EMERGENCY PLANNING ZONES)

This case involves finding the shortest and the least cost path for Hydrogen Sulphide pipe line by constructing the proper models needed. Spatial risk score matrix is being developed. Different and needed zones are determined. These obtained data are fed to GIS system to find the targeted spatial model for the study case.

A. Objective

The goal is to implement spatial risk model derived from H2S (Hydrogen Sulphide) dispersion model to be the foundation of ERP management in GIS (Geographic Information System).

B. Scope

The scope is to create dispersion risk model for H2S derived from EUB (Energy and Utility Board) dispersion calculation model. Then, calculated EPZ (Emergency Planning Zones) spatially for sour wells based on phase operation is implemented. Wind magnitude and direction impact in case of calculated protective action zone are considered..

Capability of creating actual EPZ zones is based on geospatial analysis relationships. These zones are recognized with respect to the available geographic objects. For example, these objects can be the access routes and their availability..

Implement the spatial risk assessment score matrix per sour well to enable the spatial management of ERP type related to the well based on its status.

C. Proposed spatial data model for ERP management based on the spatial risk model

Figure 1 shows the proposed spatial ERP data model based on the developed spatial risk model derived from the defined spatial and non-spatial probable parameters as well as the related consequence analysis for H2S dispersion through which the emergency level could be specified as well as related ERP modules. Different factors related to the spatial situation in question is being taking care off. Figure 2 shows proposed flow chart of spatial ERP management based on EPZ consequence analytical zones.

D. Customized spatial risk model using python script in ESRI environment

The data needed are being gathered. These data are being put in the proper form for ESRI computer software environment. Figure 3 shows the customized GIS model used to extract the main output parameters from the calculated EUB (Energy and Utility Board) for H2S calculation model. Such parameters are the calculated EPZ (Emergency Planning Zone), PAZ (Protective Action Zone), IIZ (Initial Isolation Zone) distances, H2S concentration, wind magnitude, phase operation, and other mandatory parameters to be accounted for in the spatial risk model.

E. Create spatial risk model

This model will indicate the amount of risk involved depending on the probability analysis needed. The spatial EPZ, PAZ, IIZ per analytical asset or element (sour well) is being created. Then, the spatial risk model can be obtained based on EPZ consequence analysis and the defined impact parameters.

Figure 4 shows the initial output of the spatial risk H2S dispersion model based on the calculated EPZ zones from EUB for H2S dispersion model, as well as the creation of initial spatial risk model.

F. The final resized EPZ, PAZ, IIZ with the modified spatial risk model

The final spatial model will detect and adapt to changes. Figure 5 shows how the spatial model is intelligent enough to detect any change in the defined risk parameters whether spatial, environmental (e.g., wind direction and magnitude from sensors, etc.) or non spatial risk parameters (e.g., uncontrolled flow, etc.) and adapt accordingly.

III. ILLUSTRATIVE EXAMPLE FOR SPATIAL DATA MODEL OF ELECTRIC POWER NETWORK

This case presents how to find the shortest and the least cost path for electrical line by obtaining the proper models needed. The spatial location ID code consists of the following needed codes: Utility code - Province code - City code - District or village code - Distributer code.

A single line diagram shown in Figure 6 is composed of five medium voltage cells. It shows two incoming cells, two outgoing cells, and one bus tie cell. Each cell is controlled and protected by one medium voltage circuit breaker and two isolating switches.

The spatial location ID shown in TABLE I provides the operator with the illustrative specific location of each element of the concerned network starting from the electric distribution company to the concerned element i.e., distribution board, in each cell, etc.

Each of the geographic area or sites or department belonging to the concerned electric network are shown in TABLE II provided with the code of each of them.

IV. EQUATIONS FOR SPATIAL RISK SCORE MATRIX DEVELOPMENT AND ERP SPATIAL MANAGEMENT

In this section, the needed and used mathematical formulae are being presented. The zones EPZ, PAZ, IIZ are created for each case study based on the evaluated consequence analytical values:

1) *Refined EPZ radius = EPZ calculated (1 + Related consequence value).* [1] [2].

The actual EPZ is calculated based on the impact of spatial and non-spatial parameters on the consequence matrix.

2) *(Total Of Risk) TOR = (Probability * Consequence) per asset or analytical element.* [1] [2].

The total of risk score per asset(object). This value is inserted into the risk score matrix for risk evaluation.

3) $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$. [4]

Binomial distribution equation: that represents the probability relationship between the quality performance and related risk. This equation is used for the calculation of quality performance indicator of the target plan.

where $\mu = \frac{TOR}{2}$,

and $\sigma = \sqrt{\frac{TOR^2}{4}}$

X = 1 (Assumed Random Variable)

Scale Factor = 100 (Assumed)

4) *QPI (Quality Performance Index)* $= f(x) \times 100$

QPI of target plan is calculated assuming that the scale factor is 100.

V. CONCLUSION

In this paper, the role of spatial model is being applied to obtain integration between various workflows. The concept of VCDM (Virtual Common Data Model) concept is used. The SDM (Spatial Data Model) is created to act as the common data model based on analyzing the activities and workgroups. Spatial analysis for the decomposed workgroups within common processes is proved to be more manageable and scalable than relying on common elements in the development of the domain spatial data model. The obtained spatial model is dynamic and robust.

VI. REFERENCES

[1] "British Columbia Oil and Gas Handbook Emergency Planning and Requirements for Sour Wells,". Technical guideline documentation from BC Oil and Gas Commission.

[2] The Canadian Association of Petroleum Producers (CAPP), "CAPP Companion Planning Guide to ERCB Directive 071,". Technical guideline documentation from the Canadian Association of Petroleum Producers (CAPP).

[3] S. P. S. Raghav, and J. K. Sinha, "Electrical Network Mapping and Consumer Indexing using GIS," Map India 2006.

[4] A. Sami, "Role Of Geographic Information System For Asset Management Information Risk Assessment For Pipelines Utility In Oil And Gas Industry," ESRI, UC, July 2012.

[5] K. Johnson, "GIS Energy Management For University Of Redlands", ESRI, UC, July 2003.

[6] B. R. Williams, "Leveraging GIS For Utility Enterprise Information Management", GITA, Annual Conference, 2007.

[7] H. T. Hassan, and M. F. Akhtar, "Mapping Of Power Distribution Network Using Geographical Information System", International Journal of Emerging Technology and Advanced Engineering, vol. 2, issue 6, June 2012.

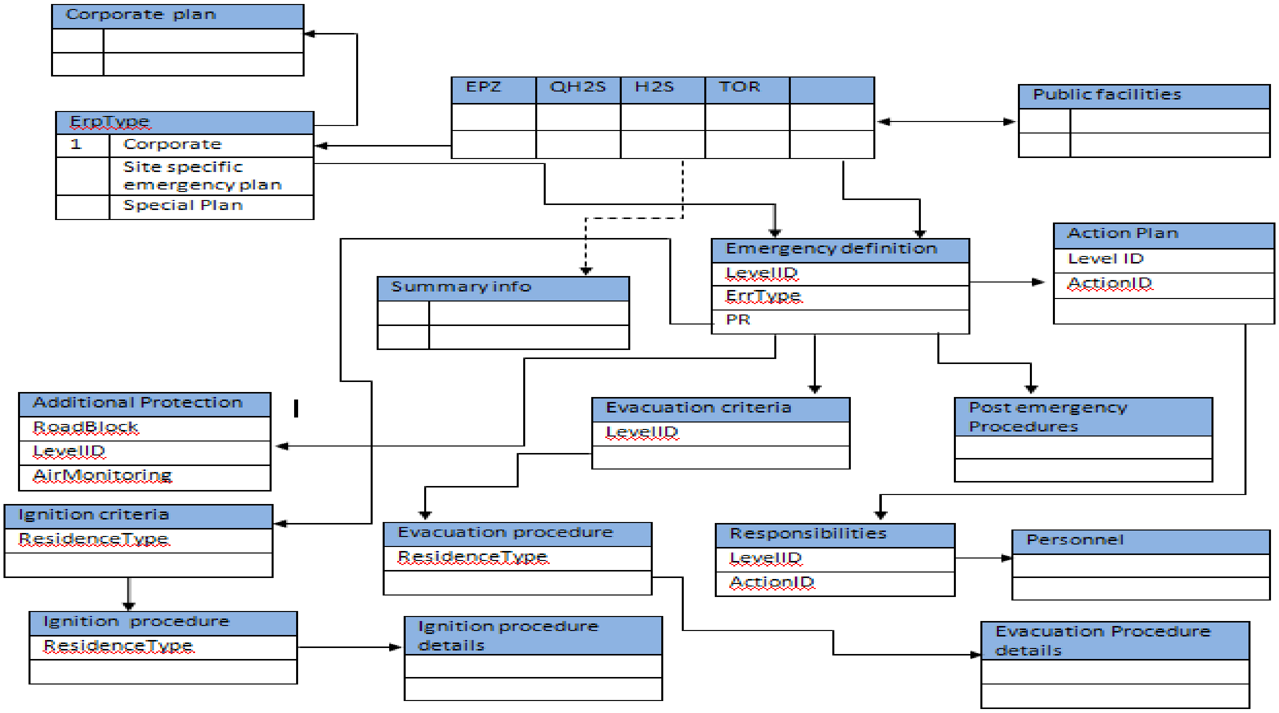


Figure 1. Example of spatial ERP data model in GIS environment.

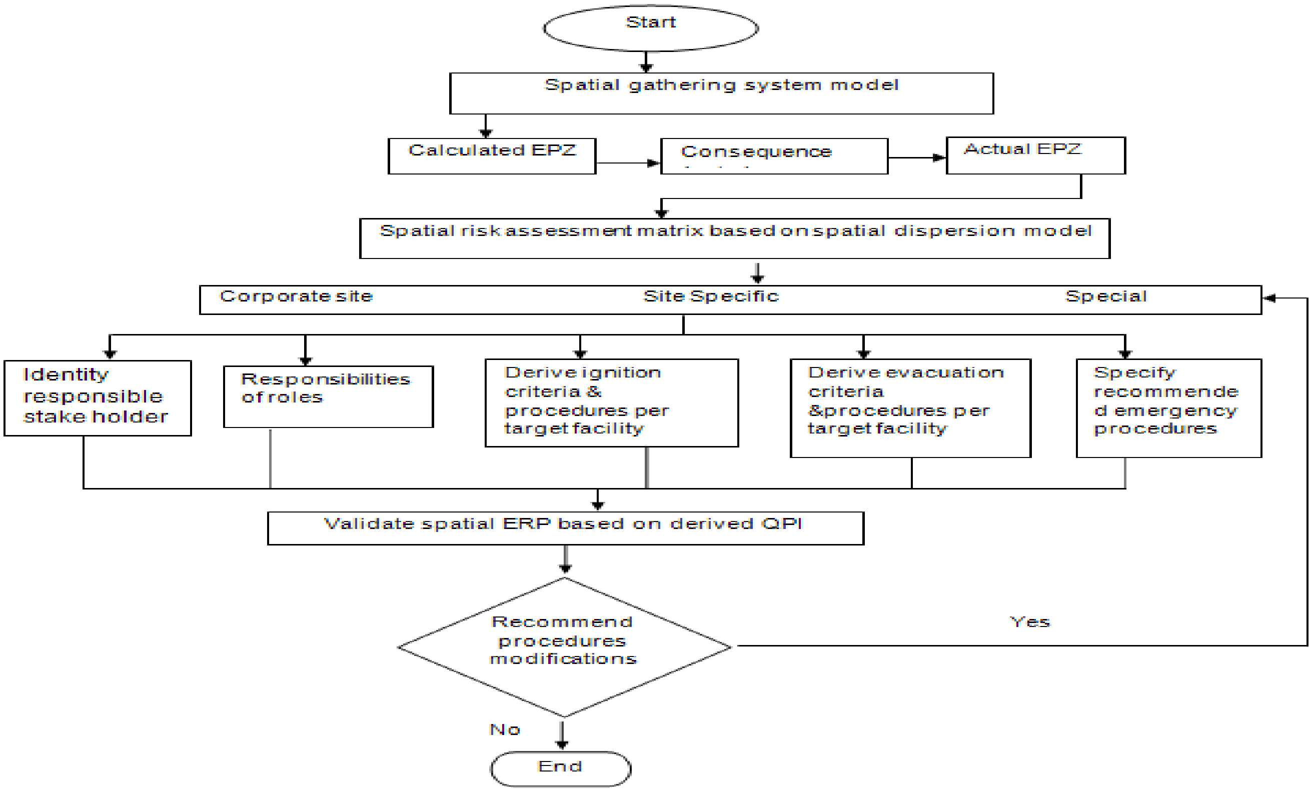
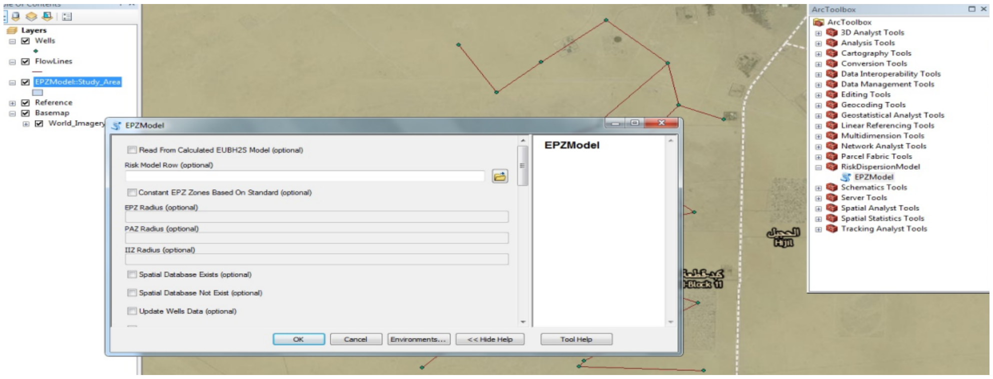

Start
Spatial gathering system model
Calculated EPZ
Consequence
Actual EPZ
Spatial risk assessment matrix based on spatial dispersion model
Corporate site
Site Specific
Special
Identity responsible stake holder
Responsibilities of roles
Derive ignition criteria & procedures per target facility
Derive evacuation criteria &procedures per target facility
Specify recommended emergency procedures
Validate spatial ERP based on derived QPI
Recommend procedures modifications
Yes
No
End

Figure 2. Proposed flow chart of spatial ERP management based on EPZ consequence analytical zones.



EPZModel

□ Read From Calculated EUBH2S Model (optional)

Risk Model Row (optional)

□ Constant EPZ Zones Based On Standard (optional)

EPZ Radius (optional)

PAZ Radius (optional)

IIZ Radius (optional)

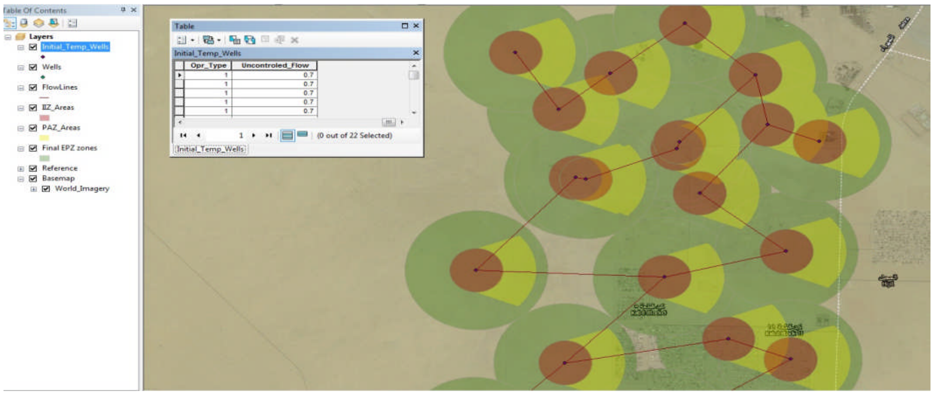
□ Spatial Database Exists (optional)

□ Spatial Database Not Exist (optional)

□ Update Wells Data (optional)

EPZModel

Figure 3. The model interface of spatial H2S dispersion model.



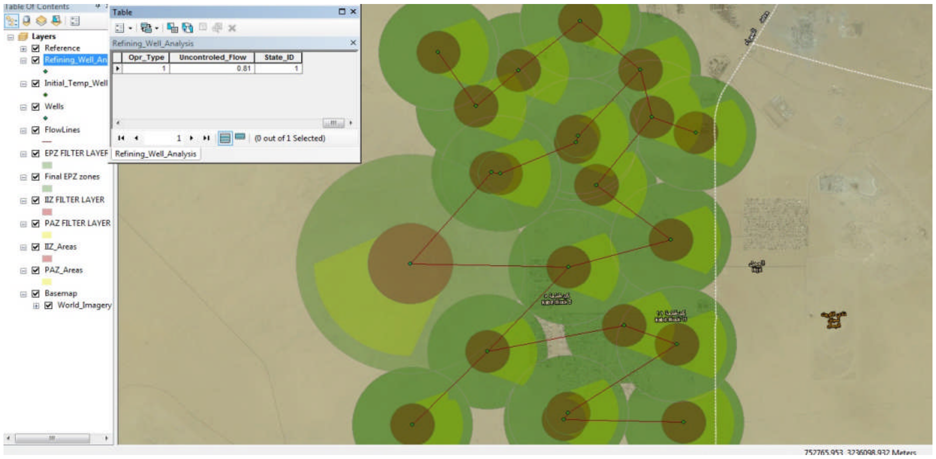
Initial_Temp_Wells

	Opr_Type	Uncontroled_Flow
▸	1	0.7
	1	0.7
	1	0.7
	1	0.7
	1	0.7

(0 out of 22 Selected)

Initial_Temp_Wells

Figure 4. The calculated spatial EPZ zones.



Refining_Well_Analysis

	Opr_Type	Uncontroled_Flow	State_ID
▸	1	0.81	1

(0 out of 1 Selected)

Refining_Well_Analysis

Figure 5. The refined spatial EPZ zones.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-383-4

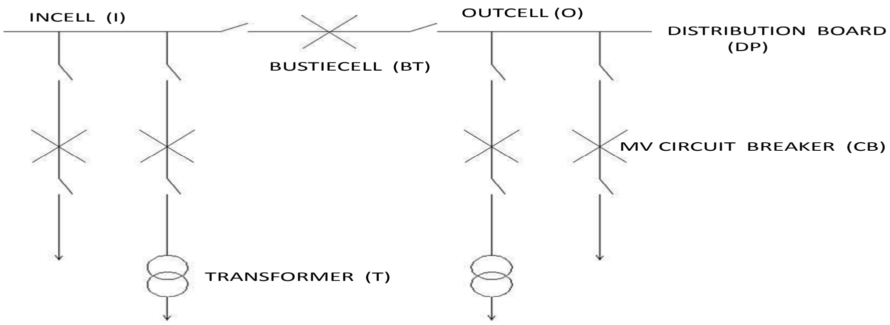


Figure 6. Single line diagram of Distribution Board (DB)

TABLE I. SPATIAL LOCATION ID CODE OF ELECTRIC POWER NETWORK.

No.	Item Name	Location ID
1	Distribution Board (11.0 & 22.0 KV)	6/1/02/DP05
1.1	In Cell	6/1/02/DP05/I1
1.1.1	MV Circuit Breaker (11.0 or 22.0 KV)	6/1/02/DP05/I1/CB1
1.2	Out Cell	6/1/02/DP05/O1
1.2.1	MV Circuit Breaker (11.0 or 22.0 KV)	6/1/02/DP05/O1/CB1
1.3	Bus Tie Cell	6/1/02/DP05/BT1
1.3.1	MV Circuit Breaker (11.0 or 22.0 KV)	6/1/02/DP05/BT1/CB1

TABLE II. GEOGRAPHIC ELEMENTS CODE.

Name	*Code*
Electricity Distribution company	6
governorate	1
district	02

ISBN: 978-1-61208-383-4

Temporally Transformed Map Animation for Visual Data Analysis

Salla Multimäki and Paula Ahonen-Rainio
Dept. of Real Estate, Planning, and Geoinformatics
Aalto University, School of Engineering
Finland
e-mail: salla.multimaki@aalto.fi

***Abstract*— Animated maps are widely used in visualizing the temporal aspect of geographical data, even though their effectiveness depends on multiple factors and is far from obvious. This paper presents one method to complement map animations with a transformation that manipulates the temporal dimension. This temporal transformation, in which events were located equally into a time period, applies the idea of spatial transformation, which is familiar from cartography. We performed a user test with a transformed animation of two different datasets to demonstrate how the transformation presents different spatio-temporal relationships. The test indicates that the transformed animation can reveal spatio-temporal patterns which cannot be detected from a traditional animation, such as the order of very dense events. The users found the transformed animations insightful and useful. For complex visual analysis the combination of the transformed animation and timeline visualization seems to be the most effective.**

Keywords- map animation; temporal transformation; user testing.

I. INTRODUCTION

A map animation is a common method for visualizing spatio-temporal information. The reason for this is simple: an animated map allows spatial information to be presented on a map and, at the same time, it naturally uses the temporal dimension for presenting changes over time. However, an animated map does not automatically result in effective comprehension. The comparison between animations and static visualization methods has been considered in many studies, both in graphics [1] [2] and in cartography [3] [4] [5], and the results from these studies show variation in the superiority of the methods, depending on the types of tasks and datasets.

Fabrikant et al. [6] point out that well-designed map animations are inherently different from well-designed static maps and comparison between these two methods is not even meaningful. They argue that instead the attention should be focused on studying when and why these methods do work. Lobben [7] showed that animations are better suited when users' tasks deal with time, and also found some evidence that static maps could work better with location-based tasks. Simultaneous changes at many discrete locations are difficult to perceive from an animation but as a presentation of general spatio-temporal patterns and the behaviour as a whole an animation can be most valuable [8].

It is evident that user control affects the usability of an animation and the effectiveness of its comprehension. In many cases, it improves users' performance [9], but at the same time, user control may distort the continuity of the animation so much that the advantages of the animation are lost [8]. Especially if an animation contains long still periods and/or dense temporal clusters, the ability to control the speed of the animation becomes essential. However, it is clear that the user control tools always increase the users' cognitive load and produce a split attention problem [10].

These inefficiencies suggest that in order to support the interpretation of map animations, we should search for methods that overcome the limitations of animations. These kinds of methods would be especially required in exploratory tasks when making sense of spatio-temporal behaviour. With static maps, cartographic transformations such as density transformation are used to present a phenomenon from different perspectives and to reveal patterns that would otherwise stay hidden. This inspired us to study the possibility of benefiting from similar transformations of the temporal dimension in animated maps. We performed a concept testing with interviews with the test users to test the effectiveness of temporal transformation in detecting spatio-temporal patterns.

The theoretical background and practical details of the transformation are explained in the next two sections. After that, the test animations, test setting, and interviews, as well as the results, are presented. Finally, discussion of the results takes place and conclusions are drawn.

II. MOTIVATION

The spatial and temporal dimensions of geographic data share commonalities, such as scale and its relation to the level of detail of the phenomena that are represented [11]. When we present a real-world space on a map, we shrink the presentation into a smaller scale and, in most cases, explicitly inform the users about the scale. In a map animation, real-world time is usually correspondingly scaled into a shorter display time. Despite the fact that the temporal scale, just like the spatial scale, has a strong influence on the observation and understanding of the phenomena, the temporal scale is not commonly calculated and expressed numerically. Instead, the passing of time is presented as a relative location of a pointer on a time slider.

There are also similarities between the temporal and spatial relationships of objects. The temporal topological relationships presented by Allen [12] show many similarities to spatial data: moments in time can be ordered, and temporal objects can be equal to, meet, overlap with, or

ISBN: 978-1-61208-383-4

include each other. However, because of the one-dimensional nature of time, the temporal order is unambiguous and each point object can have only two neighbours in time: the one which is the closest before and the one after. The only temporal metric relationship is the length of time (the duration of an event or an interval between two events), corresponding to distance in space.

The duration, order, and rate of change of the dynamic visual variables [13] make it possible to present different spatio-temporal datasets meaningfully in an animation. Typically, the duration of the scenes is kept constant throughout the whole animation, and in animations which present change in time, the order of the scenes is chronological. The rate of change depends on the sampling rate of the real-world phenomenon in the dataset. The duration and rate of change together produce the perceived speed of the animation. If we want the change in the animation to look smooth, the rate of change must be small enough and successive scenes must follow each other fast enough. If the sampling rate is too low, extra data frames should be interpolated between them to give the impression of smooth animation.

The idea of many of the geographic transformations, such as generalization and equal density transformation, makes sense when they are adapted to the time dimension, while some other transformations cannot strictly be applied to the time axis because of the one-dimensional nature of time. Monmonier [14] gives examples of how to apply spatial generalization operators such as displacement, smoothing, and exaggeration to the temporal dimension of dynamic statistical maps, but his motivation is to avoid incoherent flicker and twinkling dots and instead allow the perception of salient patterns in dynamic thematic maps. Kraak [15] has also experimented with the transformation that he calls "from time to geography", with static presentation. It stretches the famous Minard's Map according to the temporal dimension in such a way that the time periods with slow or no movement stretch the spatial representation of the trajectory. This example differs from so-called travel time cartograms since it does not consider any location as a reference point to which the travelling times are calculated. Andrienko et al. [16] present a time transformation called a "trajectory wall", where the time axis of a space-time cube is modified to present the relative order of the events. Therefore, trajectories that follow the same route do not overlap with each other, but form a wall in which the order of the trajectories on the time axis is determined by their starting times.

In this study, we propose the temporal transformation of animation in order to present the data from a different perspective and with a different temporal emphasis in order to support the analysis of spatio-temporal phenomena. Our research question was whether this transformation can be comprehensible and useful to users. For the study, we created transformed map animations of two different datasets with different temporal structures. As a result of the transformation, the events in a dataset were evenly located in time over the whole time period, keeping their order. The basis of this transformation is somehow similar to a trajectory wall, since it preserves the order of the events, but it is implemented with a dynamic display. We call this transformation "equal density transformation" because it equalizes the time intervals between consecutive events.

III. EQUAL DENSITY TRANSFORMATION

Spatially, in equalized density transformation the areas of high density of the phenomenon are made bigger and the areas of low density become smaller, so that the spatial density of the phenomenon becomes constant [17]. This transformation is presented in an example in Figure 1. The distances between points are equalized and the reference grid stretches and shrinks correspondingly. In the temporal version of equalized density transformation, the time intervals between each two consecutive events are equalized in length over the whole time period. Equal density transformation is performed by dividing the events evenly into the time period of an animation. If the dataset contains events that feature duration, their start and end times are simply handled as separate objects on the timeline. In this transformation, the accurate timestamps of events are lost, but the temporal topological relationships remain constant. Should there be any events with exactly the same timestamp, their mutual order must be determined by some other attribute.

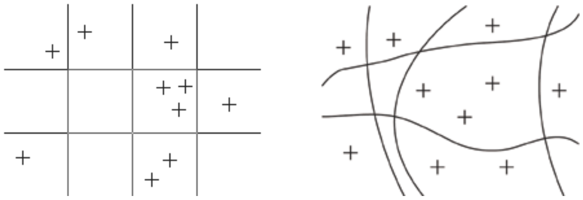


Figure 1. Spatial equalized density transformation illustrated with a small example dataset.

Two example sets of events in their original form and after the equal density transformation are shown in Figures 2a and 2b. From them it can be seen that the density of the events reflects the degree of the transformation. When the events are condensed, their time period is stretched to last longer. Consequently, time periods with sparse events are "fast-forwarded". This reduces the user's temptation to fast-forward those periods when some potentially important information may remain unobserved.

IV. USER TEST AND INTERVIEWS

In this section, we first describe the animated dataset, the making of the test animations, and the implementation of the test and user interviews. The results from the test and interviews are presented in the next section.

A. Test Data and Animations

The aims of the user test were to find out whether the test users prefer the transformed animation or the original, whether they find the transformation useful, and whether they really understand the effect of the transformation. Therefore we prepared a set of questions that the test users were to answer while viewing the animations. The number of times they viewed each animation and the additional comments they made during the test were recorded.

The dataset used in this test contained Twitter messages, so-called tweets, from the area of Port-au-Prince, Haiti, from a four-month period after the earthquake in January 2010. Twitter was used to search for help and food or water supplies and also to find missing persons. A Twitter user can allow the exact coordinates of the tweets to be saved and shown by the service provider, and all these tweets were included in the test dataset.

For this test, two different datasets were prepared. The first dataset covered the four-month period after the earthquake, but to keep the size of the dataset reasonable, every tenth tweet was selected. In this dataset, most of the tweets were strongly compressed into the first days and weeks of the time period, and after that the density of the tweets decreased remarkably. This dataset is henceforth referred to as the “Every 10th” dataset and is shown on a timeline in Figure 2a. The other dataset contained the very first tweets right after the earthquake. To achieve the same number of objects (193 tweets), the dataset was cut to cover about an 84-hour period. Because of the problems in electricity production in Haiti, only those tweets that were sent between 6 am and 6 pm were successfully published. This caused strong periodicity in the data. This dataset is henceforth referred to as the “First days” dataset and is shown in Figure 2b. Both datasets were equal density transformed by using Microsoft Excel. In the transformed datasets, the time interval between two consecutive events is the whole time period divided by the number of events. Figure 2c shows the effect of the equal density transformation; these two datasets become similar. The timestamps are not visible in this timeline, because they were artificially modified and did not correspond to the real-world time.

Four map animations were made with ArcGIS10; two presented the original “Every 10th” and “First days” datasets and two presented the equal density transformed datasets. All animations were of equal length, 60 seconds, and they each contained 193 events. The events were presented on a background map with red dots that appeared brightly and faded to a less saturated red after that. A screenshot from one animation is shown in Figure 3. In addition to the animations, the timeline visualizations were presented in the test view immediately below the test animations to help the test users to comprehend the temporal patterns of the data.

B. *Test Setting*

At the beginning of the test, the main concepts of the test, such as pattern and spatio-temporal information, were introduced to the user and the user was able to familiarize himself/herself with an example of the animations, layout, and arrangements of the user interface.

The test contained two parts. One part presented the original and transformed animations of the “First days” dataset and the other part the corresponding animations of the “Every 10th” dataset. Each test user performed both parts, but the order of these parts varied between the users in order to avoid the influence of the learning effect. These two parts were identical in terms of their layout and arrangements.

In the first phase, the user first had the opportunity to view only the original animation as many times as he/she wanted, and after that was asked to answer Questions 1.1 and 1.2 (Table 1). The questions dealt with the overall impression of the dataset. Then the user viewed the temporally transformed animation and answered the same questions. The order of the animations was fixed to this, because we wanted to simulate the explorative analysis task where the user first gets an overview of the data and then uses more complex tools, focusing on more detailed analysis.

In the second phase, the user could use both of these two animations to answer three more detailed questions (Questions 2.1-2.3 in Table I) about the behavioural patterns of the data. Because of the differences in the datasets, the questions varied slightly between the two datasets.

After finishing both parts of the test, the user was interviewed. The interview was semi-structured and covered the following topics:

- Was the temporal transformation as a method easy to understand?
- What could have made the transformation easier to understand?
- Did you use the animation, timeline, or still picture of the animation to answer the questions?
- Was the transformed animation useful when answering the questions? Why?
- In what tasks was the transformation especially useful?
- Could this kind of tool be useful in your job?

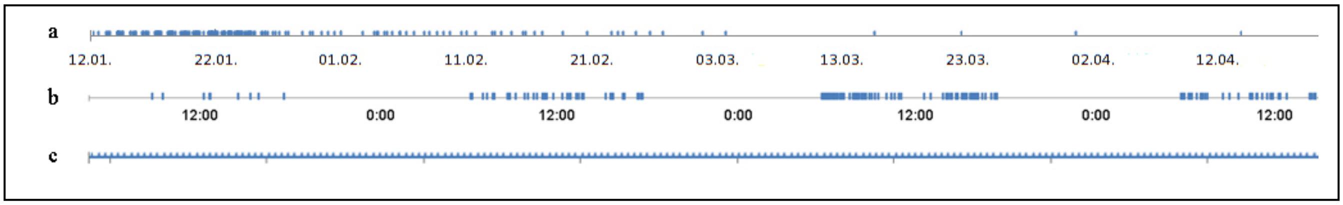


Figure 2. The top row (a) shows the “Every 10th” dataset visualized on a timeline. The middle row (b) shows the “First days” dataset visualized on a timeline. The scales of the timelines are different. The bottom row (c) shows both datasets after the equal density transformation.

Some users had already discussed these topics during the test, and in these cases not all the questions were explicitly gone through during the interview.

TABLE I. QUESTIONS IN THE USER TEST.

Q	“First days” dataset	“Every 10th” dataset
1.1	What kind of patterns do you find from the data?	
1.2	Are there any events which seem not to fit the data or draw your attention in some other way?	
2.1	Where are the first and last events of the dataset located?	In what area are the first ten events of the dataset located?
2.2	Is there a location on the map where there are multiple sequential events? Where is it?	Are there time periods when the events are clustered into a certain area?
2.3	Is there an area on the map where the events are clustered both spatially and temporally?	Does the centroid of the events move during the animation?

The test was completed by nine users. They were professional cartographers or geographers with experience of temporal datasets. Four of them were female and five male. Their ages varied between 28 and 55 years.

The users did the test on a laptop computer that was connected to a data projector. The evaluator observed the user’s performance via the data projector, calculating the viewing times for each animation. The users could answer the test questions either by typing their answers into a textbox on the display or verbally to the evaluator, who wrote those answers down. The answers in the interviews were also written down by the evaluator.

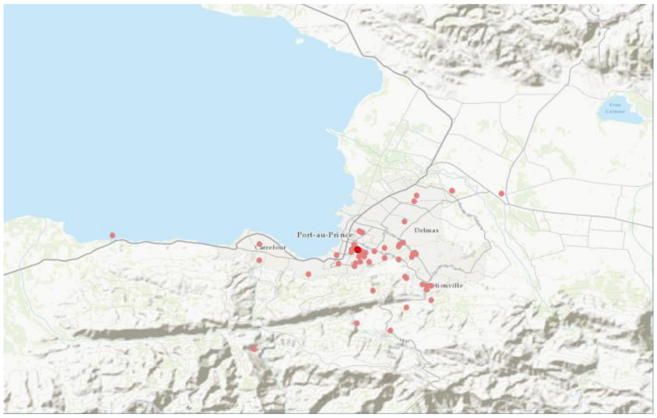


Figure 3. A screenshot from the transformed animation of the “First Days” dataset. The latest event is seen as a bright red dot in the middle of the map. Basemap: Esri 2013.

C. Analysis of the Material

From the user test, the following indicators were analyzed:

1. the number of times the user viewed each animation in each task, and whether he/she viewed the whole animation or the viewing was discontinued;
2. the kinds of behaviour patterns the user found from the animations and whether these findings were appropriate;
3. whether the user’s impression of the phenomenon represented in the dataset differed between the original and transformed animations.

From the users’ comments during the test and their answers in the interview, the following factors were calculated:

1. positive and negative comments the users made about each animations;
2. faulty and inaccurate interpretations that the users made from the animations;
3. cases where the user made different interpretations from the same dataset on the basis of the two animations.

Because of the small number of test users, no statistical significance parameters were calculated from these results.

V. RESULTS

In the test, the users chose to view the transformed animation slightly more often than the original animation. This trend was particularly clear with Questions 2.1 and 2.2, which dealt with so-called elementary lookup tasks [18]. With more complex analysis tasks, the users tended to interrupt the flow of the original animations by pausing or fast-forwarding, while the transformed animations were more often viewed in their entirety. This pattern can be seen in Table II. The first, **bold** number in each box marks the times when the animation was viewed completely, and the second number (in parentheses) marks the times when the animation was viewed partially, which means that the user paused, fast-forwarded, or interrupted it in some other way during the viewing. Every row in the table corresponds to one task in the test.

The differences between the results for the two datasets in Question 2.3 are caused by the difference in the questions. With the “First Days” dataset, the question concerned the whole time period, and therefore the users had no choice but to view the animation completely. On the contrary, with the “Every 10th” dataset the task was to find a spatio-temporal cluster, and the users could stop viewing the animation after finding the first one.

TABLE II. THE VIEWING TIMES OF EACH ANIMATION

	Every 10th dataset		**First Days dataset**	
	original	***temporally transf.***	***original***	***temporally transf.***
Q 1.1 and 1.2	**16**(2)		**16**(2)	
Q 1.1 and 1.2		**15**(2)		**17**(0)
Q 2.1	**2**(10)	**3**(14)	**5**(6)	**3**(12)
Q 2.2	**4**(5)	**9**(2)	**6**(1)	**11**(0)
Q 2.3	**6**(10)	**9**(4)	**9**(1)	**8**(0)

ISBN: 978-1-61208-383-4

When viewing the transformed animation of the “First days” dataset, in the first phase of the test six out of the nine users mentioned that they perceived a location at which several events appeared sequentially. This location can be seen as the latest and brightest dot in Figure 3. Later, when explicitly asked (Q2.2), the remaining three users also perceived it. From the original animation, none of the users perceived this kind of behaviour at first, and after the question Q2.2, only one user mentioned that he also saw that phenomenon in the original animation, “but much more weakly than in the transformed animation”. This location with sequent events proved to be a police station in the centre of Port-au-Prince. Our assumption is that the inhabitants of the city went to the police station to seek their missing relatives after the earthquake, and the police used Twitter to support the search and rescue efforts.

The users learned to favour the transformed animation with some tasks even during this kind of very short test. For the questions Q2.1 and Q2.2 the transformed animation was used approximately 40% more often than the original, and was viewed more often without interruption. With the question Q2.3 the preference between the animations varied, depending on the dataset, but the transformed animation was viewed in its entirety more often.

In the interviews most of the users had a positive attitude towards the transformed animations. They were apparently pleased and said that the transformation was “charming” or “nicer”. They also mentioned that the transformed animation was “better” and “easier to watch”. Two users said that the transformed animation was “exhausting” and its “continuous info flow was tiring”. However, these two users also commented that the transformation was useful in some cases. A summary of elements calculated from the interviews is shown in Table III.

TABLE III. SUMMARY FROM THE INTERVIEWS

	Original animation	**Temporally transformed animation**
More useful (pos.)	2	14
Unpleasant to view (neg.)	1	2
Misinterpretations		2
Different interpretations	8	

The users had varying opinions about the applicability of the transformation. For some tasks performed in the test, some of the users said that the transformed version was “essential”, while others said that it did not suit those tasks. When asked about the use cases for this kind of transformation, the users mentioned several possible application areas in addition to elementary lookup tasks dealing with time. For example, traffic planning, crowd movement analysis, environmental analysis, and oil destruction activities were proposed. The users also pointed out the possibility of combining the analysis of the temporal dimension on the timeline with behavioural analysis of the transformed animation.

From the interviews it became clear that a proper temporal legend could have improved the performance; five of the nine test users mentioned this when asked about development ideas. More specifically, the idea of colouring the events according to their timestamp was mentioned by several users. Another suggestion was to improve the linking between the timeline and the animation; a moving pointer should show the flow of time of the transformed animation on the timeline of the original data.

The test results indicate that it is essential to ensure that the user understands how the transformation influences the animation. In several cases (eight out of the 27 behaviour descriptions recorded) the users’ impression about the phenomenon varied between the original and transformed animations, even though they knew that the animations presented the same dataset. Some clear misinterpretations appeared; in one case the user grouped the last events of one day and the first events of the next day into the same spatio-temporal cluster despite the fact that there was a 12-hour gap. The same user also made a false statement about the location of the last event of the dataset.

VI. DISCUSSION AND FURTHER RESEARCH

The user test shows that the equal density transformation of the animation revealed the position of sequential events that were not detected from the original animation. We believe that this pattern would have remained unnoticed without the transformation. Additionally, this pattern was found spontaneously; in the majority of cases it arrested the attention of the users without any specific search task. This is an important feature of the visual analysis tool when it is used for data exploration and data mining.

As the interviews with the users proved, the power of the equal density transformation lies in the fact that it can reduce the user’s need to interrupt the animation, and therefore offers a smooth overall evocation of the phenomenon. At the same time it eases the cognitive load on the user by offering a continuous, temporally predictable change with no congested periods. It emphasizes the order of the events and equalizes them in relation to time, thus attributing equal significance to all the events.

The findings indicate that the disadvantage of the transformation is that misinterpretations of the transformation are possible or even probable. These misinterpretations could be minimized with a temporal legend in which the user can always perceive the phase of the animation. Furthermore, properly designed colours for the events could indicate the degree of the transformation, and these colours could be used for linking the timeline and the animations.

The idea of colouring the events according to their timestamp also arose repeatedly in the interviews. The users suggested, for example, that the colour of the point-type events might change smoothly day by day. This would help the user to detect spatio-temporal clusters from the map and also communicate about the temporal discontinuities in the data.

However, a disadvantage of the use of colour as a temporal legend is that one cannot visualize any attribute data by means of colours at the same time. When the events are point-type and visualized with small, round objects, other ways to present attribute information are limited. Therefore, consideration should be given to whether the combination of these two variables with the use of colour is possible. Brewer [19] proposed a set of colour scheme types to be used with bivariate data, but this set does not contain the combination of qualitative attribute information and a bipolar subordinate variable. The temporal transformation can affect the time either by stretching or shrinking it, and therefore its visualization should also be bipolar. If the degree of temporal transformation is simplified to binary data (= slower or faster than the original), then Brewer's "qualitative/binary" combination can be used. In this schema, the qualitative data is visualized with different hues and the binary data is visualized with the lightness of the colour. It must be noted that Brewer's model is designed for choropleth maps, and its applicability to point-type data is not obvious. Therefore it is clear that more research is needed to test whether discrimination between these two variables is possible in a use case similar to the case in this study.

Another possible way to provide the information about the speed of the animation is sound. [20] suggests sonic input to represent the passing of time. This could also be a useful technique with an animation with changing speed, since human hearing is relatively sensitive to changes in rhythm and pitch. A user test with a sonic legend and a larger number of test users is our future research plan.

The lack of attribute information is a limitation of this study. In this study, we tested the concept of equal density transformation and therefore wanted to keep the test arrangement as simple as possible. It would be possible, for example, to classify the tweet messages according to their thematic content, and visualize this classification by the use of colour.

Because of the simplification of the test procedure, the user control over the animations was also limited. The users did not have a chance to adjust the speed of the animation or to filter its content. However, we offered the most common user control tools; playing, pausing, and the opportunity to jump to any moment in the animation. A wider selection of control tools would have increased the cognitive load on the user and drawn the user's attention away from the task being tested.

VII. CONCLUSIONS

The human ability to adopt information from temporal animation is limited. If the animation runs too fast, is too long, or presents too many events simultaneously, a user can easily miss some information, and therefore is not able to form a full image of the phenomenon being presented. The traditional control tools of an animation, such as pausing, jumping to a specified scene, or looping, have a limited capability to improve this understanding.

This paper presented a novel method for transforming the temporal dimension of map animations by equalizing the interval between two consecutive events. The user test proved that the transformation can reveal patterns that would have been left unnoticed with traditional animation. It seems to be understandable for the users and useful for spatio-temporal analysis in parallel to an original, non-transformed animation.

In exploratory analysis a rich variety of tools that complement each other is a necessity. The results from this user test and interviews indicate that equal density transformation might be an appropriate technique to complement a set of such analysis tools.

REFERENCES

[1] B. Tversky, J. B. Morrison, and M. Betrancourt, "Animation: can it facilitate?" International Journal of Human-Computer Studies, vol 57, 2002, pp. 247-262.

[2] D. Archambault, H. C. Purchase, and B. Pinaud, "Animation, small multiples, and the effect of mental map preservation in dynamic graphs," IEEE Transactions on Visualization and Computer Graphics, vol. 17(4), 2011, pp. 539-552.

[3] A. Koussoulakou and M.-J. Kraak, "Spatio-temporal maps and cartographic communication," The Cartographic Journal, vol. 29(2), 1992, pp. 101-108.

[4] T. S. Slocum, R. S. Sluter, F. C. Kessler, and S. C. Yoder, "A qualitative evaluation of MapTime, a program for exploring spatiotemporal point data," Cartographica, vol. 39(3), 2004, pp. 43-68.

[5] A. Griffin, A. MacEachren, F. Hardisty, E. Steiner, and B. Li, "A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters," Annals of the American Cartographers, vol. 96(4), 2006, pp. 740-753.

[6] S. I. Fabrikant, S. Rebich-Hespanha, N. Andrienko, G. Andrienko, and D. R. Montello, "Novel method to measure inference affordance in static small multiple displays representing dynamic processes," The Cartographic Journal, vol. 45(3), 2008, pp. 201-215.

[7] A. Lobben, "Influence of data properties on animated maps," Annals of the Association of American Geographers, vol. 98(3), 2008, pp. 583-603.

[8] M. Harrower and S. Fabrikant, "The role of map animation for geographic visualization." In: M. Dodge, M. McDerby, and M. Turner (Eds.), Geographic Visualization, John Wiley & Sons, 2008.

[9] Kraak, M.-J., Edsall, R., and MacEachren, A., "Cartographic animation and legends for temporal maps: Exploration and or interaction." In: Proceedings of the 18th International Cartographic Conference, Stockholm, Sweden, 1997, pp. 253-260.

[10] M. Harrower, "The cognitive limits of animated maps," Cartographica, vol. 42(4), 2007, pp. 349-357.

[11] G. Andrienko et al., “Space, time, and visual analytics,” International Journal of Geographical Information Science, vol. 24 (10), 2010, pp. 1577-1600.

[12] J. Allen, “Maintaining knowledge about temporal intervals,” Communications of the ACM, vol. 26(11), 1983, pp. 832-843.

[13] D. DiBiase, A. M. McEachren, J. Krygier, and C. Reeves, “Animation and the role of map design in scientific visualization,” Cartography and Geographic Information Systems, vol. 19(4), 1992, pp. 201-214, 265-266.

[14] M. Monmonier, “Temporal generalization for dynamic maps,” Cartography and Geographic Information Science vol. 23 (2), 1996, pp. 96-98.

[15] M.-J. Kraak, Mapping Time: Illustrated by Minard's Map of Napoleon's Russian Campaign of 1812, ESRI Press, California, 2014.

[16] G. Andrienko, N. Andrienko, H. Schumann, and C. Tominski, “Visualization of trajectory attributes in space-time cube and trajectory wall”, in M. Buchroitner et al. (eds.), Cartography from Pole to Pole, Lecture Notes in Geoinformation and Cartography, Springer-Verlag Berlin Heidelberg, 2014, pp. 157-163.

[17] B. FitzGerald, Science in Geography 1: Developments in Geographical Method. Oxford University Press, 1974.

[18] N. Andrienko and G. Andrienko, Exploratory Analysis of Spatial and Temporal Data – A Systematic Approach, Springer, Berlin, Germany, 2006.

[19] C. A. Brewer, “Guidelines for use of the perceptual dimensions of color for mapping and visualization”, IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics, 1994.

[20] M. Kraak, R. Edsall, and A. M. MacEachren, “Cartographic animation and legends for temporal maps: Exploration and or interaction”, Proceedings of the 18th International Cartographic Conference, 1997, pp. 253-260.

On Maxmin Active Range Problem for Weighted Consistent Dynamic Map Labeling

Xiao Zhang
Department of Computer Science,
City University of Hong Kong
Hong Kong
Email: xiao.zhang@my.cityu.edu.hk

Sheung-Hung Poon
Department of Computer Science,
Institute of Information Systems
and Applications, National Tsing Hua University
Hsin-Chu, Taiwan
Email: spoon@cs.nthu.edu.tw

Minming Li, Victor C.S.Lee
Department of Computer Science,
City University of Hong Kong
Hong Kong
Email: {minming.li, csvlee}@cityu.edu.hk

Abstract*—Geographical visualization systems, such as online maps, provide interactive operations on continuous zooming and panning. In consistent map labeling, users can navigate continuously through space without distracting behaviors such as popping and jumping. We study the consistent dynamic map labeling problem: Given a set of labels on the map and each label with a selectable active range and weight, find an appropriate active range for each label such that no two consistent labels intersect at any scale and the minimum weighted active range is maximized. It is named as *maximizing minimum weighted active range problem* (*MMWAR*). This study on MMWAR is of the theoretical and practical significance, since it is common that some labels in practical maps need better visibility than others. We investigate both the *simple* and *general* variants and present several theoretical results. For simple variants, simple 1D-MMWAR and 2D-MMWAR with proportional dilation are optimally solved in $O(n \log n)$ and $O(n^2 \log n)$, respectively. For general variants, we prove that general 1D-MMWAR with constant dilation and 2D-MMWAR with proportional dilation are NP-complete. Moreover, we provide an $O(\log n)$-approximation algorithm for the general 1D-MMWAR with proportional dilation, and an $O(\sqrt{n})$-factor approximation algorithm for the general 2D-MMWAR with proportional dilation. Our experimentation results show that on average, the approximation factors in our algorithms are much smaller than the worst-case upper bounds stated above, and our approximation algorithms run efficiently.

Keywords–Geographic information systems; Dynamic map labeling; NP-hardness; Approximation algorithms.

I. Introduction

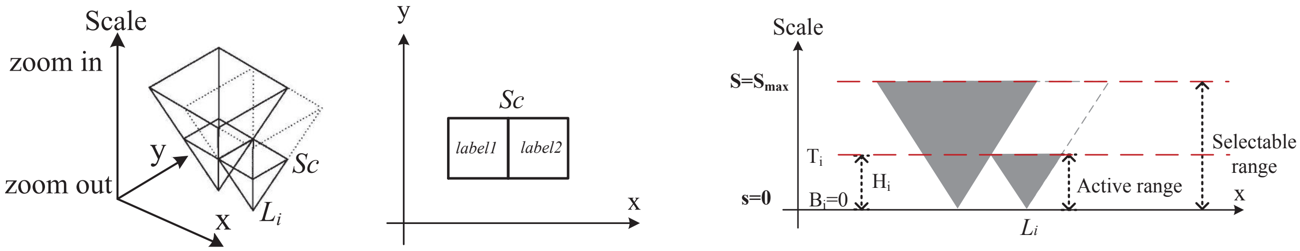
Dynamic map labeling, as a critical problem in cartography and geographic information systems (GIS), provides users interactive operations on zooming and panning maps continually and dynamically. In contrast to the static map labeling problem [1], it can be formulated as a traditional map labeling problem by incorporating *scale* as an additional dimension. Increasing academic concern is aroused to handle these interfaces. Been et al. [2] initially defined the consistent dynamic map labeling problem with a set of consistency desiderata to provide a new and practical framework. According to this framework, during zooming and panning, (a) labels are not allowed to exhibit abrupt change in the position or size; (b) labels should not suddenly disappear and reappear when zooming in or pop up when zooming out; (c) the labeling should be in line with the selected map viewpoint, not be hinged on the navigation history.

Most previous algorithmic studies on consistent dynamic map labeling deal with active range optimization (ARO) problem [3][4], whose objective is to maximize the sum of total active ranges, each of which corresponds to the consistent interval of scales with visible labels. On the other hand, maximizing minimum active range problem is seldom considered, since a few labels may only have a very small selectable range [5]. Nevertheless, the maximizing minimum weighted active range problem, arises in a natural way but with practical importance in situations when, different cities may have different weights in order to reveal the different degrees of importance. For example, on a map of China, attributing Beijing a higher priority (weight) than Tianjin (a nearby city of Beijing) ensures that in case of limited space the capital rather than one of its nearby cities receives a label. Clearly this maximizing minimum weighted active range (MMWAR) problem is equivalent to finding a set of active ranges with different visibility that the overall weighted range assignment is relatively balanced. In particular, none of the existing dynamic map labeling methods provides theoretical studies and related solutions on MMWAR.

In this paper, we study the problem MMWAR and propose a suite of algorithms. The present paper is structured as follows. Section II describes some preliminary concepts. The related work is presented in Section III. The complexity of MMWAR is investigated in Section IV. An algorithmic study of MMWAR is presented in Section V. We give exact algorithms for simple 1D-MMWAR and 2D-MMWAR with proportional dilation. The general 1D-MMWAR and 2D-MMWAR with proportional dilation are provided with approximation algorithms, whose performance is evaluated by experiments in Section VI. Section VII concludes this paper and discusses some open problems.

II. Preliminaries

In this study, we adopt the model of consistent dynamic map labeling and all above mentioned desiderata to our problem [2]. Each label is represented by a three dimensional (3D) solid. It is formed by extruding the label shape through the vertical dimension (*zooming scale*). Each solid can be truncated to a single scale interval, named its active range (or height, for short), corresponding to the scale selected by the label. The labels are assumed not to slide and rotate. See Figure 1, we consider invariant point placements with axis-aligned square labels. The output of our proposed algorithm is a set of disjoint active ranges. See Figure 1(a), at any zooming scale Sc, we obtain a set of disjoint labels at the cross section at scale Sc. This set of labels represents the labeling of the points we considered at this specific scale Sc. See Figure 1(b) as an example. In the dynamic setting of zooming in and out



(a) Front view of square frusta (b) Placement of labels at scale Sc (c) Side view of square labels

Figure 1. Two square labels with selectable ranges and active ranges

of the scene, different labeling at different zooming scales of the solution will be shown to the users.

We give several notations on the MMWAR problem. In principle, the input is a set of extrusions, each extrusion with a selectable interval and weight, for which we intend to assign an active range. Assume that we are given a set of extrusions $\Gamma = (L_1, L_2, ..., L_n)$, each L_i with a selectable range (s_i, S_i) and weight w_i, where $s_i, S_i \geq 0$ are at most the global maximum scale $S_{\max}$. Our objective is to calculate an active range assignment A for Γ. More specifically, for each extrusion $L_i \in \Gamma$, a contiguous active range $(B_i, T_i) \subseteq (s_i, S_i)$ is assigned. $H_i = T_i - B_i$ indicates the height of active range. A is a feasible solution if the resulting extrusions are pairwise disjoint. For a solution A, we say that an extrusion L_i is active in A at scale Sc if $Sc \in (B_i, T_i)$; otherwise L_i is inactive in A at scale Sc. The goal is to maximize the minimum weighted height of A, i.e., $MaxMin\frac{H_i}{w_i}$.

Following the work [5], we consider two variants in this problem, i.e., *simple* and *general*. The problem stated in the previous paragraph is the *general* variant. For the simple variant, it is of importance to theory and practice to consider the scenario in which all labels are selectable at all scales and all labels are selected when zooming in. Specifically, in simple MMWAR variant, $(s_i, S_i) = (0, S_{\max})$, $B_i = 0$ for all $L_i \in \Gamma$; see Figure 1(c). Moreover, we study two kinds of dilation cases in this paper, i.e., *proportional dilation* and *constant dilation*. We say that labels have proportional dilation if their sizes could change with scale proportionally. In contrast, if the sizes of labels are fixed at every scale, we say that labels have constant dilation. For the simple MMWAR problem with proportional dilation, the shapes of extrusions appear to be rectangular pyramids by extruding rectangular labels. For the general MMWAR problem with proportional dilation, the shapes of extrusions appear to be congruent square pyramids or frusta segments of congruent square cones.

III. Related work

Map labeling is essential for a wide range of applications and becomes the focus of research [6][7]. Before the proposal of dynamic labeling problem, there was a large number of studies on automated label placement dealing with *static* fixed position [5]. The research outputs covered various settings and NP-hardness proofs [1][8]. A typical task of these works was to select and place labels without intersection so as to maximize the number of selected labels. Exact and approximation algorithms [9] [10][11] are known for several types of the static label optimization problem. Among them, Agarwal et al. [12] proposed a PTAS for the unit-width rectangular label placement problem and a $\log n$-approximation algorithm for the arbitrary rectangle case. Then, the improvement was obtained in [10]. Chalermsook and Chuzhoy [11] studied the *Maximum Independent Set of Rectangles* problem and presented an $O(\log \log n)$-approximation algorithm.

In recent years, dynamic map labeling has become a new bright spot. Petzold et al. [13] generated a reactive conflict graph storing all potential conflicts information by using preprocessing phase. Poon and Shin [14] presented algorithms for labeling points that precomputed a hierarchical data structure to store solutions at different scales. For consistent dynamic map labeling, Been et al. [2] initially presented several consistency desiderata and formulated a new algorithmic framework for fast and consistent labeling. They also showed the NP-hardness of optimal active range selection problem. In addition, several approximation algorithms were given for 1D and 2D labeling problems [3]. Nöllenburg et al. [15] explored three extensions of the one-sided boundary labeling model allowing continuous zooming and panning. Moreover, Gemsa et al. [16] gave an FPTAS for the continuous sliding model of the ARO problem. Yap [17] summarized a few open ARO problems in dynamic map labeling. More recently, Liao Chung-Shou et al. [4] considered the dynamic map labeling problem with a set of rectangles and gave several approximations to maximizing the sum of total visible ranges.

When considering the objective of maximizing the minimum active ranges, Gemsa et al. [18] not only considered *MaxTotal* ARO problem, but also studied *MaxMin* ARO problem. They showed both problems are NP-complete. However, they only considered the continuous map rotations. None of the existing research about *MaxMin* ARO studies the zooming setting. Furthermore, few researches are incorporating weights into consideration in the map labeling problem formulation. Poon et al. [19] first defined static label-placement models for labeling static points with weights and presented several corresponding algorithms. Schwartges [20] assigned labels to map objects like cities or streets and used the weights to determine the importance of a label. This is a good attempt to reduce the gap between theory and practice. Since dynamic map labeling is still an active research line, some unsolved problems remain, such as MMWAR.

IV. Complexity

In this section, we prove that two variants of MMWAR are NP-complete by reductions from the NP-hard problem *Planar 3SAT* [21]. An instance of *Planar 3SAT* is a *3SAT* formula Φ whose variable-clause $Graph_{\Phi}$ is planar.

A. General 1D-MMWAR with constant dilation

We start by considering the NP-hardness proof of the general 1D-MMWAR with constant dilation.

Theorem 1: General 1D-MMWAR with constant dilation is NP-complete; i.e., given a set of n axis-aligned rectangular extrusions $\Gamma = (L_1, L_2, ..., L_n)$ with weight w_i for each L_i and a real number $K > 0$, it is NP-complete to decide whether there is a valid assignment A of active ranges to Γ with $Min\frac{H_i}{w_i} \geq K$. The problem is still NP-complete even when all extrusions are squares of two different sizes and with the same weight.

B. General 2D-MMWAR with proportional dilation

In the general 2D variant of MMWAR, we give the NP-hardness proof by assuming that all extrusions are congruent square pyramids with two different weights. We note that for the variant with pyramids of only one weight, whether it is NP-complete is still open.

Theorem 2: General 2D-MMWAR with proportional dilation is NP-complete; i.e., given a set of n axis-aligned rectangular cones $\Gamma = (P_1, P_2, ..., P_n)$ with different weights w_i and a real number $K > 0$, it is NP-complete to decide whether there is a valid assignment A of active ranges to Γ with $Min\frac{H_i}{w_i} \geq K$. The problem is still NP-complete even all extrusions are congruent square pyramids with two different weights.

V. Triangles 1D-MMWAR & Cones 2D-MMWAR

A suite of algorithms are devised to solve several variants of 1D- and 2D-MMWAR problems. In the simple variants, the active ranges start from *Zero* scale. On the other hand, the active ranges start from any scale in general version, which is closer to the reality with practical significance.

A. Simple 1D-MMWAR

In simple 1D-MMWAR with proportional dilation, each extrusion is an inverted triangle with top edge attached to the horizontal line $s = S_{max}$ and apex located on the x-axis, i.e., $(B_k, T_k) \subseteq (0, S_{max})$ and $B_k = 0$, thus active range height $H_k = T_k$. The truncated extrusions differ only by heights of top edges. Figure 1(c) shows an example of active ranges assignment for the labels. Observe that the objective of the problem is $MaxMin\frac{H_i}{w_i}$ for each extrusion L_i.

Let $\Gamma = (L_1, L_2, ..., L_n)$ be the set of extrusions, and let p_i be the apex of triangle-shaped extrusion L_i on the x-axis. Assume that $p_1, ..., p_n$ are arranged from left to right. For each extrusion L_k, $k < n$, we define the left side edge and right side edge as E_k^l and E_k^r. Without loss of generality, we assume that, for every two adjacent extrusions L_k and L_t $(t > k)$, E_k^r and E_t^l intersect at h_{kt}. We denote the scale of h_{kt} as H_{kt}. They are stored as a *Doubly Linked List DuLinkList[n-1]*, in which, E_k^r and E_{k+1}^l are the left pointer field and the right pointer field, and $H_{k(k+1)}$ are the value field. Then, we construct a *RB-Tree* $\Re$ to store $(H_{12}, H_{23}, ..., H_{(n-1)n})$.

Hence, we give the exact algorithm with low time complexity, as shown in Algorithm 1.

Theorem 3: Simple 1D-MMWAR with proportional dilation can be solved in $O(n \log n)$ time.

Proof: For each pair of adjacent extrusions L_s and L_t, only one of them is assigned the active range height H_{st} in simple 1D-MMWAR. By handling intersecting points of all pairwise extrusions in Algorithm 1, none of them intersect after range assignments.

Algorithm 1 Compute the maximum minimum weighted active range for simple 1D-MMWAR with proportional dilation

```
Input: $\Gamma = (L_1, L_2, ..., L_n)$, a selectable range $(0, S_{max})$
and weight $\pi = (w_1, w_2, ..., w_n)$
Output: $MaxMin\frac{H}{w}$, HList
for each extrusion $L_i$ in $\Gamma$ do
    DuLinkList $\leftarrow L_i$
end for
Construct RB-Tree $\Re$ from DuLinkList
for the minimum element $H_{ij} \in \Re$ do
    find two corresponding extrusions $(L_i, L_j)$ with weights
    $(w_i, w_j)$
    if $\frac{H_{ij}}{w_i} \geq \frac{H_{ij}}{w_j}$ then
        add $L_i$ with $H_{ij}$ to the list HList
        delete $L_i$ and update DuLinkList and $\Re$
    else
        add $L_j$ with $H_{ij}$ to the list HList
        delete $L_j$ and update DuLinkList and $\Re$
    end if
end for
$MaxMin\frac{H}{w} \leftarrow min\{\frac{H}{w}\}, \{H \in HList, w \in \pi\}$
Return $MaxMin\frac{H}{w}$
```

Optimality. For the smallest $H_{\alpha\beta}$ in the RB-Tree, it denotes the lowest intersecting point of two extrusions L_α and L_β with weights w_α and w_β. The value $H_{\alpha\beta}$ must be assigned to either L_α or L_β. Assume that $w_\alpha > w_\beta$, thus $\frac{H_{\alpha\beta}}{w_\alpha} < \frac{H_{\alpha\beta}}{w_\beta}$ (weighted active ranges). It shows that $\frac{H_{\alpha\beta}}{w_\alpha}$ is the smaller one. The only way to increase $\frac{H_{\alpha\beta}}{w_\alpha}$ is to assign $H_{\alpha\beta}$ to L_β and remove it to a candidate set $HList$, since the remaining values of *H* are larger than $H_{\alpha\beta}$. Thus, the weighted active range of L_α can be increased to be larger than $\frac{H_{\alpha\beta}}{w_\alpha}$. That is to say, whenever we select one of two intersecting extrusions, we select the one whose weighted active range is larger and add it to the candidate set $HList$. By this means, all the smaller weighted active ranges can be maximized and restored to $HList$.

Complexity. For the running time, we construct the *RB-Tree* by using $T_1(n)$. Thus, we have

$$T_1(n) = O(\log n!) < O(n \log n).$$

For each operation on the *RB-Tree*, the running time $T_2(n) = O(\log n)$. Hence, the overall time complexity of Algorithm 1 follows.

$$T(n) = T_1(n) + T_2(n) = O(n \log n).$$

Therefore, we can obtain the optimal solution of $MaxMin\frac{H}{w}$ in $O(n \log n)$ time. ■

B. Simple 2D-MMWAR

The idea of Algorithm 1 for 1D-MMWAR can be easily extended to solve the simple 2D-MMWAR.

In contrast to 1D-MMWAR, we need to construct a RB-Tree to store all pairs of the lowest intersection of 3D cones, whose amount is $O(n^2)$. Similar to Theorem 3, each time we choose the smallest H_s from the RB-tree for comparing the

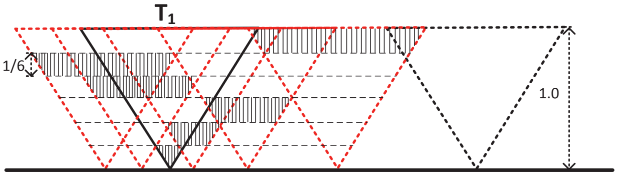


Figure 2. An illustration of the intersection degree and active ranges assignment.

weighted active range $\frac{H_s}{w}$ of two intersecting cones. Then, we remove those cones involving with the larger weighted active range from the RB-Tree. There are $O(n)$ 3D cones, which implies we need to delete those H_s from RB-tree in total $O(n \log n)$ time. Hence, the time complexity is $O(n^2 \log n)$.

Theorem 4: Simple 2D-MMWAR with proportional dilation can be solved in $O(n^2 \log n)$ time.

C. General 1D-MMWAR

In this subsection, we give an approximation algorithm that yields $\log n$-approximation for the general 1D-MMWAR with proportional dilation.

Suppose that we are given a set of congruent triangles or trapezoidal segments of congruent triangles with apexes on the x-axis. They are assumed to be with the same weight. As described in Algorithm 2, the idea is to choose the triangle intersecting with the maximum number of other triangles, then assign the active ranges to these triangles as evenly as possible. In addition, we need a new piece of notation Δ, i.e., *intersection degree* of triangle T, which denotes the number of triangles that intersect with triangle T in the graph. As shown in Figure 2, there are *eight* congruent triangles, whose selectable ranges are assume to be *one* unit, i.e., $S_{max} = 1$. If triangle T_1 is intersecting with $five$ triangles (red triangles), we say that the intersection degree of triangle T_1 is $five$, which is denoted as $\Delta = 5$.

See Algorithm 2 for the pseudo-code of our algorithm. In the following theorem, we show that such a solution is in fact an $O(\log n)$-approximation for the general 1D-MMWAR problem with the same weight.

Algorithm 2 Compute the maximum minimum weighted active range for general 1D-MMWAR with proportional dilation

```
Input: a set $\Gamma$ of $n$ congruent triangles, a selectable range $(0, S_{max})$ and weight $w$
Output: $MaxMin\frac{H}{w}$
for each triangle $T_i \in \Gamma$ do
    $\Delta_i \leftarrow$ the number of intersecting triangles
end for
$\Delta_{max} \leftarrow Max\{\Delta_1, \cdots, \Delta_n\}$
for each triangle $T_i \in \Gamma$ do
    $H \leftarrow \frac{S_{max}}{\Delta_{max}+1}$
end for
Return $\frac{H}{w}$
```

Theorem 5: Given a set of n congruent triangles with the same weight, a $\log n$-approximation for the general 1D-MMWAR with proportional dilation of *General Congruent Triangles* can be computed in $O(n^2)$ time.

Proof: Given n congruent triangles with the same weight, whose selectable ranges are assumed to be *one* unit, i.e., $S_{max} = 1$. Considering the triangle T_m intersecting with the maximum number of triangles Δ_{max}, Let X_T be the set containing T_m and all the triangles intersecting with T_m. Thus, $|X_T| = \Delta_{max} + 1$. According to Algorithm 2, each triangle $T \in X_T$ is assigned an active range height $\frac{1}{\Delta_{max}+1}$. Obviously, none of the active ranges of the triangles in X_T conflicts with each other after range assignment, since they are assigned evenly. Then, considering the case that triangle $T^{'} \notin X_T$ with intersection degree Δ', observe that $\Delta' \leq \Delta_{max}$. $X_{T'}$ contains T' and all the triangles intersecting with T', in which we assume there are k common triangles in $X_{T'}$ and X_T, $k \leq \Delta'$. Observing that each of those k common triangles has been assigned an active range of height $\frac{1}{\Delta_{max}+1}$, which occupy a range of total height $\frac{k}{\Delta_{max}+1}$. For the remaining $\Delta' + 1 - k$ triangles, they have the total range of height $\frac{\Delta_{max}+1-k}{\Delta_{max}+1}$ to be assigned. We assign each triangle an active range of height $\frac{1}{\Delta_{max}+1}$ for these remaining $\Delta' + 1 - k$ triangles. Since $\Delta' \leq \Delta_{max}$, all triangles in $X_{T'}$ and X_T can be assigned the active ranges without conflict. Note that when $k = 0$, $X_{T'}$ and X_T are disjoint subsets of triangles.

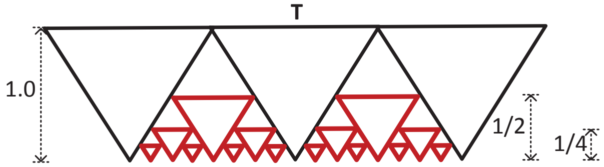


Figure 3. An illustration of finding the optimal solution for general 1D-MMWAR with proportional dilation.

When $n = 1$, the active range is S_{max}. When $n = 2$ or 3, the solutions are $\frac{S_{max}}{2}$ and $\frac{S_{max}}{3}$, respectively. As for any case that X_T with maximum intersecting degree contains m $(3 < m \leq n)$ triangles, we consider the optimal solution S^* for m triangles in this case. The solution is illustrated as *bold red triangles* in Figure 3. The gaps between disjoint triangles can be extracted up to $m - 3$ triangles in $O(m \log m)$ time. Then, we calculate the summation of all selected active range $A(t)$ for each $t \in X_T$ in the solution from the bottom to top, i.e.,

$$\sum_{t \in X_T} A(t) = 3 + 2 \times \frac{1}{2} + 2^2 \times \frac{1}{2^2} + 2^3 \times \frac{1}{2^3} + \ldots + 2^i \times \frac{1}{2^i}.$$

With the purpose of reaching the amount of triangles m, we have

$$i = \log(m-1) - 1 \Rightarrow \sum_{t \in X_T} A(t) = 2 + \log(m-1)$$

It indicates that the optimal solution follows.

$$S^* \leq \frac{\sum_{t \in X_T} A(t)}{m} = \frac{2 + \log(m-1)}{m}$$

In this case, we obtain the solution $S = \frac{1}{m}$.

Hence, Algorithm 2 achieves the approximation factor $\frac{S^*}{S} = 2 + \log(m-1) \leq 2 + \log(n-1)$. Thus, we obtain an $O(\log n)$-approximation for the general 1D-MMWAR with proportional dilation. ■

We can extend our algorithm to handle triangles with two weights. Hence, we obtain the following Corollary 1.

Corollary 1: Given n congruent triangles with the two weights w_1 and w_2 such that w_1 is a constant factor of w_2, Algorithm 2 computes an $O(\log n)$-factor approximation for the general 1D-MMWAR with proportional dilation.

D. General 2D-MMWAR

Suppose that we are given a set of congruent square pyramids or frusta segments of congruent square cones with apexes on the horizontal plane. For simplicity, we set $S_{max} = 1$. Our algorithm, say Algorithm 3, for the general 2D-MMWAR runs greedily using the same method as Algorithm 2. The only difference is that we are now considering congruent square frustra instead of congruent triangles. The pseudo-code of new algorithm is given in Algorithm 3.

Algorithm 3 Compute the maximum minimum weighted active range for general 2D-MMWAR with proportional dilation

```
Input: a set $\Gamma$ of $n$ congruent frusta, a selectable range $(0, S_{max})$ and weight $w$
Output: $MaxMin\frac{H}{w}$
for each frustum $T_i \in \Gamma$ do
    $\Delta_i \leftarrow$ the number of intersecting frusta
end for
$\Delta_{max} \leftarrow Max\{\Delta_1, \cdots, \Delta_n\}$
for each frustum $T_i \in \Gamma$ do
    $H \leftarrow \frac{S_{max}}{\Delta_{max}+1}$
end for
Return $\frac{H}{w}$
```

Theorem 6: Given a set of n axis-aligned congruent square frusta with the same weight, an $O(\sqrt{n})$-approximation algorithm for the maximum minimum weighted active range can be computed in $O(n^2)$ time.

Proof: The correctness proof uses the similar approach as the proof in Theorem 5. Recall the notation of intersection degree Δ, each extrusion is assigned only one active range of height $\frac{1}{\Delta_{max}+1}$, none of which intersect with each other after range assignments. So, we start by considering the cone T intersecting with the maximum number of cones, denoted as X_T. We show that the cone T is intersecting with other eight cones to formulate a *square* in 2D plane. For i-th cutting procedure, the amount of the cones reaches $(2^i + 1)^2$. As illustrated in Figure 4, when $i = 2$, we extract one cone with height $\frac{1}{2}$ from each pairwise disjoint cones with height 1. Thus, when the amount of the cones reaches the number n, we obtain that $i = \log(\sqrt{n} - 1)$. Then, we calculate the summation of all selected active range $A(t)$ for X_T in the solution from bottom to top.

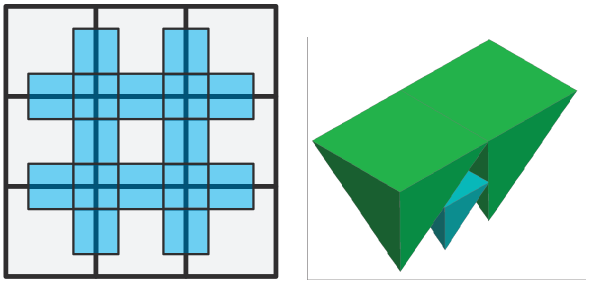
$$\sum_{t \in X_T} A(t) = 3\sqrt{n} + 2\log(\sqrt{n} - 1) - 2$$

which implies the optimal solution S^* as follows.

$$S^* \leq \frac{\sum_{t \in X_T} A(t)}{n} = \frac{3\sqrt{n} + 2\log(\sqrt{n} - 1) - 2}{n}$$

Furthermore, Algorithm 3 gives a solution $\frac{1}{\Delta_{max}+1} \geq \frac{1}{n}$. Overall, Algorithm 3 achieves the approximation factor $\frac{S^*}{S} \leq 3\sqrt{n} + 2\log(\sqrt{n} - 1) - 2$. Thus, we obtain an $O(\sqrt{n})$-approximation for the general 2D-MMWAR with proportional dilation. ■

We can extend our algorithm to handle congruent square frusta with two weights. Hence, we obtain the following Corollary 2.



(a) An example of the top-down view.

(b) 3D model of cutting the gap between each pairwise disjoint cones.

Figure 4. Illustration of the cutting procedure for general 2D-MMWAR with proportional dilation

Corollary 2: Given n congruent frusta with two weights w_1 and w_2 such that w_1 is a constant factor of w_2, an $O(\sqrt{n})$-approximation for the general 2D-MMWAR with proportional dilation can be computed.

VI. Experiments and Evaluation

In this section, we evaluate the performance of Algorithm 2 for general 1D-MMWAR and Algorithm 3 for general 2D-MMWAR. Since these problems are proved to be NP-hard, we compare the results obtained by the proposed algorithms with the theoretical bounds. The experiments are conducted on a 3.4GHz Intel PC with 4GB RAM. The programming language is MATLAB(R2013a).

A. Approximation ratio for general 1D-MMWAR

For Algorithm 2, we uniformly distributed 1,000 congruent triangles along a straight-line segment. For each triangle set of size $n = 50, 100, ..., 1,000$, we randomly generate 1,000 cases. The average performance ratio is recorded in Figure 5, where the horizontal axis represents input size of congruent triangles and the vertical axis represents the average approximation ratio.

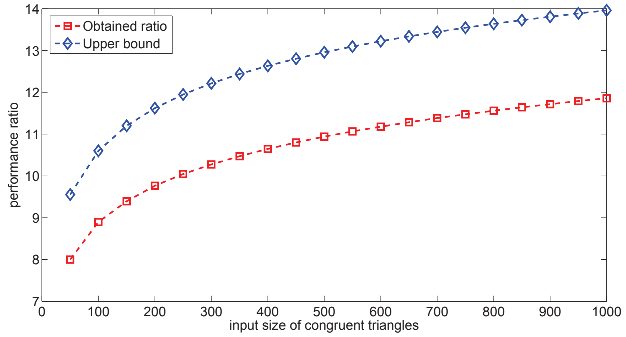


Figure 5. Comparing the approximation ratio of Algorithm 2 with its theoretical upper bound.

B. Approximation ratio for general 2D-MMWAR

For Algorithm 3, we uniformly distributed 1,000 congruent frusta in the unit square. For each frustum set of size $n = 50, 100, ..., 1,000$, we randomly generate 1,000 cases, and record the average approximation ratio in Figure 6.

Summarizing and evaluating our results, we have observed that the proposed approximation algorithms have much smaller approximation ratios than the worst-case theoretical upper bounds. Besides, it seems that, as the problem scale increases, the real approximation ratio increases little.

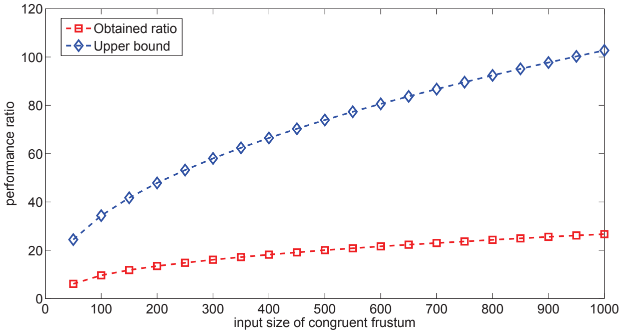


Figure 6. Comparing the approximation ratio of Algorithm 3 with its theoretical upper bound.

C. Running time for Algorithms 2 and 3

For the running time, we averaged the running time of the 1,000 cases on both algorithms with input size from 50 to 1,000 and showed the results in Figure 7. It indicates that the running times of the approximation algorithms follow the theoretical complexity bounds, and both algorithms run efficiently.

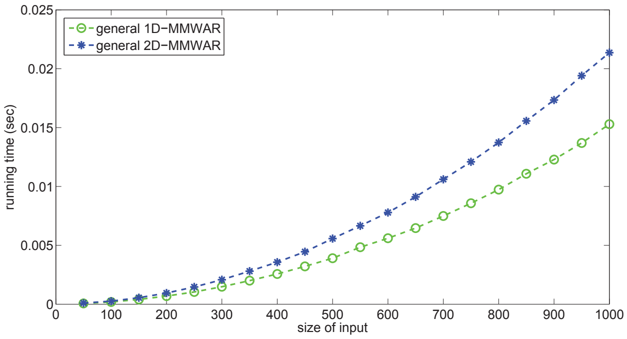


Figure 7. Average running time in seconds

VII. Conclusion and Discussion

The weighted active range optimization problem is of great theoretical and practical importance in map labeling, and this is the first work with the objective of maximizing the minimum weighted active range. We prove that general 1D-MMWAR with constant dilation and general 2D-MMWAR with proportional dilation are NP-complete. We have proposed two exact algorithms for simple 1D-MMWAR and 2D-MMWAR with proportional dilation and two approximation algorithms for general 1D-MMWAR and 2D-MMWAR with proportional dilation. For the complexity analysis, there are still several open problems. The complexity of general 2D-MMWAR with constant dilation is still unknown. For proportional dilation, since we assume that the input extrusions have two different weights in the NP-hardness proof of general 2D-MMWAR, the complexity of the problem with only one weight remains as an open problem. Furthermore, we believe that the approximation factor and time complexity of the approximation algorithms and corollaries for general 1D-MMWAR and 2D-MMWAR with proportional dilation could be further improved.

Acknowledgment

This work was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CityU 122512] and grant NSC 100-2628-E-007-020-MY3, Taiwan.

References

[1] S. Doddi, M. V. Marathe, A. Mirzaian, B. M. Moret, and B. Zhu, “Map labeling and its generalizations,” in Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 1997, pp. 148–157.

[2] K. Been, E. Daiches, and C. Yap, “Dynamic map labeling,” IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 5, 2006, pp. 773–780.

[3] K. Been, M. Nöllenburg, S.-H. Poon, and A. Wolff, “Optimizing active ranges for consistent dynamic map labeling,” in Proceedings of the 24th Annual Symposium on Computational Geometry. ACM, 2008, pp. 10–19.

[4] C.-S. Liao, C.-W. Liang, and S.-H. Poon, “Approximation algorithms on consistent dynamic map labeling,” in Frontiers in Algorithmics, ser. Lecture Notes in Computer Science, J. Chen, J. Hopcroft, and J. Wang, Eds., vol. 8497. Springer International Publishing, 2014, pp. 170–181.

[5] K. Been, M. Nöllenburg, S.-H. Poon, and A. Wolff, “Optimizing active ranges for consistent dynamic map labeling,” Computational Geometry, vol. 43, no. 3, 2010, pp. 312–328.

[6] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling, “Newsstand: a new view on news,” in Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2008, p. 18.

[7] Q. Zhang and L. Harrie, “Real-time map labeling for personal navigation,” in Proceedings of 12th International Conference on Geoinformatics-Geospatial Information Research: Bridging the Pacific and Atlantic, 2004, pp. 39–46.

[8] M. Formann and F. Wagner, “A packing problem with applications to lettering of maps,” in Proceedings of the 7th annual Symposium on Computational Geometry. ACM, 1991, pp. 281–288.

[9] G. W. Klau and P. Mutzel, “Optimal labeling of point features in rectangular labeling models,” Mathematical Programming, vol. 94, no. 2-3, 2003, pp. 435–458.

[10] T. M. Chan, “Polynomial-time approximation schemes for packing and piercing fat objects,” Journal of Algorithms, vol. 46, no. 2, 2003, pp. 178–189.

[11] P. Chalermsook and J. Chuzhoy, “Maximum independent set of rectangles,” in Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2009, pp. 892–901.

[12] P. K. Agarwal, M. Van Kreveld, and S. Suri, “Label placement by maximum independent set in rectangles,” Computational Geometry, vol. 11, no. 3, 1998, pp. 209–218.

[13] I. Petzold, G. Gröger, and L. Plümer, “Fast screen map labeling - data structures and algorithms,” in Proceedings of the 23rd International Cartographic Conference, 2003, pp. 288–298.

[14] S.-H. Poon and C.-S. Shin, “Adaptive zooming in point set labeling,” in Proceedings of the 15th International Symposium on Fundamentals of Computation Theory. Springer, 2005, pp. 233–244.

[15] M. Nöllenburg, V. Polishchuk, and M. Sysikaski, “Dynamic one-sided boundary labeling,” in Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2010, pp. 310–319.

[16] A. Gemsa, M. Nöllenburg, and I. Rutter, “Sliding labels for dynamic point labeling,” in Proceedings of the 23rd Canadian Conference on Computational Geometry, 2011.

[17] C. K. Yap, “Open problems in dynamic map labeling,” in Proceedings of 20th International Workshop on Combinatoral Algorithms, 2009, pp. 1–3.

[18] A. Gemsa, M. Nöllenburg, and I. Rutter, “Consistent labeling of rotating maps,” in Algorithms and Data Structures. Springer, 2011, pp. 451–462.

[19] S.-H. Poon, C.-S. Shin, T. Strijk, T. Uno, and A. Wolff, “Labeling points with weights,” Algorithmica, vol. 38, no. 2, 2004, pp. 341–362.

[20] N. Schwartges, “Dynamic label placement in practice,” in Proceedings of the 2nd Association of Geographic Information Laboratories for Europe PhD School, 2013.

[21] D. Lichtenstein, “Planar formulae and their uses,” SIAM journal on computing, vol. 11, no. 2, 1982, pp. 329–343.

Water Quality Impact Assessment of Land Use and Land Cover Changes

A dynamic IT model for territorial integrated management

Maria J. Vale
Directorate-General for Territorial Development
Ministry for Environment, Spatial Planning and Energy
Lisbon, Portugal
Email: mvale@dgterritorio.pt

Bruno M. Meneses
Directorate-General for Territorial Development
Ministry for Environment, Spatial Planning and Energy
Lisbon, Portugal
Email: bmeneses@dgterritorio.pt

Rui Reis
Directorate-General for Territorial Development
Ministry for Environment, Spatial Planning and Energy
Lisbon, Portugal
Email: rui.reis@dgterritorio.pt

Raquel Saraiva
Ministry for Environment, Spatial Planning and Energy
Directorate-General for Territorial Development
Lisbon, Portugal
Email: rsaraiva@dgterritorio.pt

Jesus Villegas
Grupo Tragsa
SEPI
Madrid, Spain
Email: jmev@tragsa.es

Mariano Cruz
Grupo Tragsa
SEPI
Madrid, Spain
Email: mnc@tragsa.es

***Abstract* – The land use and land cover (LULC) changes have influenced the water quality. Thus, addressing LULC can lead to the understanding of the actual problems and how to avoid future water stress problems. This paper presents a model that integrates land use and land cover changes (LUCC), and their impacts on water quality. The model is tested for the main drinking water reserve for Continental Portugal, the Castelo de Bode Dam located in the Zêzere watershed. This work integrates the specifications of SmartOpenData Project (SmOD), and is strictly related to the work developed in the eEnvPlus, and TerAGUA projects. It considers INSPIRE Directive specifications, and those related to linked open data integration. The results obtained so far address the most relevant LULC changes in Portugal, but can easily be applied to study the problem in other countries, allowing the creation of guidelines for spatial planning integrating the concerns of the Water Framework Directive.**

Keywords - LULC, LUCC; Water; Spatial Planning, SmartOpenData, INSPIRE.

I. Introduction

The land use and land cover (LULC) of the territory is constantly changing, mainly due the anthropogenic actions, but also due to natural causes. Many of these changes have a negative impact on water quality, a fact cited by some authors that show that there is a cause and effect relation between some land use/cover changes (LUCC) and the decrease in water quality [1]-[4]. The human related LUCC, namely deforestation, urbanization, urban sprawl or less adequate agricultural practices where large amounts of chemicals are applied (from industrial or domestic origin) is therefore a key factor to build well balanced land use planning initiates. Many land use practices result in an increase of contaminants that are easily transported to watercourses, contribute to the degradation of important water reserves namely those for public supply, and inducing relevant water stress [6]-[10].

Being so it is important to have an integrated watershed land use planning and management system, integrating the water legal framework with the land use planning legal framework with updated knowledge on LUCC changes and ongoing management of natural resources: water, energy production, environmental protection integrated with socioeconomic development strategies.

The interconnection between these areas is essential to understand the intensity of land use changes upstream of a drinking water reserve, in order to understand the impact of land use changes in water stress [11]. Therefore, it is essential to integrate LULC knowledge, e.g., intensive agriculture, industry location, urbanization and urban sprawl, with existing water treatment plants, the monitoring of human activities and their water management requirements. The gradual decrease in natural water quality, namely for drinking water reserves, leads to increasing pressures on water management efficiency in order to reduce migration or leaching of certain contaminants in surface and ground waters downstream. The increasing risk for public health is also relevant.

Water quality preservation is consequently a main concern for governments and for the communities as a whole. Assuring water quality for different purposes, and namely providing drinking water supply in the short and long terms, is essential to preserve the future of every Nation and is a basis for economic development [7].

Less effective approaches to management imply the need for large investments, which lead to increasing operational and maintenance costs, usually charged, directly or indirectly, to the final consumers. Thus, it is necessary to establish measures linking economic development models to efficient local, regional, national and international water management, establishing priorities to the most relevant water reserves, in line with an efficient and sustainable management of resources and management of the territory.

The provision of environmental and geographic information (linked data), in a semi-structured, shared web based platform, will be an asset for the efficiency and sustainability management models of these resources. This will allow the scientific community to monitor and further study the phenomena in order to build solutions that lead to avoid or reduce negative impacts on water quality and to assure that effective results will increase [5]-[12]-[13]. This information will be important for licensing small and medium-sized enterprises that require this natural resource, for citizens as for public institutions, so that everyone linked to the implementation of the development model can be involved [7]-[8].

The SmartOpenData Project has created a Linked Open Data Infrastructure and available data resources for biodiversity and environment protection, in relevant areas (Protected Areas and National Parks) for some European countries. This project is focused on how Linked Open Data can be applied to spatial data resource management, to public open data portals and to initiatives like GEOSS Data-CORE, Copernicus (formerly known as GMES), INSPIRE and voluntary data (OpenStreetMap, GEO Wiki, etc.), and how it can impact on the economic and sustainability progress in European Environment research and Biodiversity Protection. This will be achieved by making existing “INSPIRE based” relevant spatial data sets, services and appropriate metadata available through a new Linked Data structure. In addition, the infrastructure will provide automatic search engines that will search additional available geospatial resources (OGC and RDF structures) across the web. A RDF (Resource Description Framework) structure is used to describe the relation between two objects The point is that by re-using existing identifiers available in the Linked Open Data cloud, SmartOpenData immediately will have access to a lot of other data sources and these will be available through SPARQL queries [9]-[11].

Understanding and maintaining watershed related knowledge is a very high data consuming issue. Namely maintaining LULC updated along with many other socioeconomic relevant data repositories requires substantial financial support.

This issue could be partially addressed by similar open data approaches like the one described above or in a more structured way like the one included in [7], involving communities, enterprises and governments in a shared effort for data maintenance with a clear purpose of monitoring drinking water quality and to achieve sustainable development.

The areas with reservoirs considered strategic for the public water supply, should be integrated into the so-called "Smart Regions" crossing over protection areas like “Nature 2000 network” and requiring specific management rules, integrated in a participative web based platform, business oriented, but built upon an extranet approach so that interested citizens and public authorities can help consolidating effective LUCC changes that lead to effective drinking water preservation and supply in a cost-benefit effective way.

The proposed evaluation model for LUCC and their impacts on water quality should also contribute to the creation of sustainable measures, especially measures that lead to the smallest human pressure in forest soils, agriculture (reduction of agrochemical application), forest fires risk, and artificial unbalanced transformations, in the vicinity of the reservoirs namely near water catchments for public supply.

The relevant framework implemented so far by the researchers in this field, is in line with the concerns of the Inspire Governance Group, and can be used in future territorial planning actions namely identifying combined areas and locations that increasingly contribute to water degradation. Establishing relations of water preservation, land use identification and land use intensity allowance, is therefore essential for drinking water preservation in line with land use planning, in a cost effective way to achieve effective economic growth.

The shared information management, web based approach will furthermore increase the knowledge made available for good governance putting in line different development perspectives, from local and regional views to the national planning intents and investment; all this must be supported by an effective way of redistributing development benefits and preservation costs, taking here, as example, the well balanced watershed management, supports by a well-balanced land use planning model integrating cadaster concerns.

In Section II, we present the methodology to LUCC determinations and the relations with variations of water quality parameters. Section III presents the preliminary results and discussion of LUCC in Continental Portugal, and the relations between LUCC and water quality degradation in Castelo de Bode Dam. Section IV presents the final considerations.

II. METHODOLOGY

In Portugal, substantial work is being developed in order to build solid co-operative Web-GIS based systems to address land use and water management in an effective and equitable way. The most relevant work relates to the TerAGUA co-operative approach and is being explored in the context of the SmartOpenData Project. It contemplates the development of a prototype model that includes the evaluation of land use and land cover changes and their influenced the water quality of Zêzere watershed catchment (sub-basin of Tagus River). The datasets used in this project contemplate essentially open data (CORINE Land Cover - CLC, geomorphologic and environmental data), integrated and put in line with all relevant official data. The shared open analysis model will quantify the LUCC, its spatial distribution within the Portuguese continental area, and allow on the fly comparisons with environmental, namely water related, data. The results will be integrated in INSPIRE infrastructure, under clear descriptions and to be addressed through unique identifiers, including the INSPIRE themes, code lists, application schemas, discovery services, integrated with the RDF application (Fig. 1).

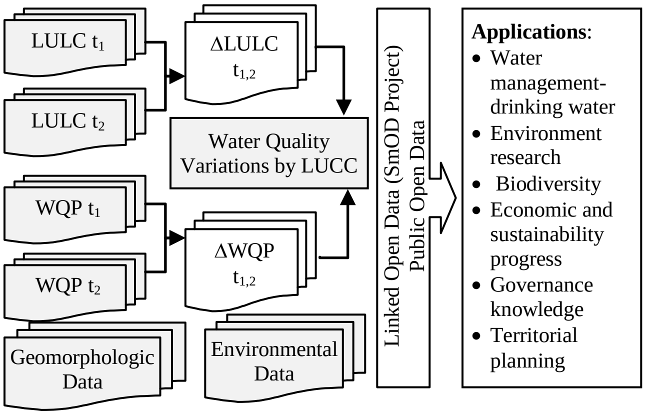


Figure 1. Model of Linked Open Data of the Portuguese Pilot (SmOD Project) and the possible aplications (WQP - Water Quality Parametrs; LULC – Land User and Land Cover; t – time).

The expected results will be used as a complement in planning actions, in particular to identify uses and locations that contribute the most to the degradation of water quality (depending on the LULC and also of LUCC that cause high negative impacts). This identification of LUCC with greater impact on surface water will be crucial to water quality preservation within the study area (economic cost reduction in line with efficient management of water treatment), but also to estimate the impact of intensively artificialized territory near water catchments, and will be used as the basis for a sustainable water and land planning in the EU.

III. Preliminary Results and Discussion

A. LUCC in Continental Portugal

In Continental Portugal the larger areas of LULC are the heterogeneous agricultural areas and forests (Table 1). According to this data, in the last two decades, forest areas have declined due to the transition to scrub and/or herbaceous vegetation, hence an increase in the area of this last type of LULC. The reduction of this type of LULC has direct implications on water quality, namely in the increase of sediments dragged to the water courses due the erosion along with the increase of chemical elements dragged or leached from exposed soils without any type of protection to storm water [2].

The registered loss of arable land may be explained to some extent by the disinvestment in agriculture and to the lack of structural support to small and medium-sized farmers, owners of small fragmented land, the most common within the Portuguese context, inducing abandonment of these lands. These soils, due to the advance of natural vegetation, began to be occupied by scrub and/or herbaceous vegetation associations, land cover that is strictly related to increasing risk of forest fire.

The area of artificial surfaces also increased, taking agricultural and forest soils. Land artificialized or built area can be highly correlated with the amount of sediment that goes into watercourses, with several aggravating factors as, for example, the amount of surface runoff waters due to soil sealing, but also the speed of runoff, causing sometimes flooding and increasing sediments transport. Increasing urban sprawl along the margins of water bodies, delivers wastewater (sewage) drainage without any domestic treatment directly to these reservoirs leading to water quality deterioration within these reservoirs.

TABLE I. LULC of Continental Portugal in Different moments. Area (%) obtained by CORINE Land Cover Data.

LULC	Year			
	1990	***2000***	***2006***	***2012***
Arable land	15,1	14,3	13,9	12,7
Artificial, non-agricultural vegetated areas	0,1	0,1	0,1	0,2
Forests	27,7	26,7	22,6	22,6
Heterogeneous agricultural areas	26,6	26,3	26,0	26,3
Industrial, commercial and transport units	0,2	0,4	0,5	0,6
Inland waters	0,5	0,6	0,8	0,9
Inland wetlands	0,0	0,0	0,0	0,0
Marine waters	0,6	0,6	0,6	0,6
Maritime wetlands	0,3	0,3	0,3	0,3
Mine, dump and construction sites	0,1	0,2	0,2	0,2
Open spaces with little or no vegetation	2,0	1,9	1,9	1,5
Pastures	0,6	0,5	0,5	0,7
Permanent crops	6,5	6,6	6,6	7,1
Scrub and/or herbaceous vegetation associations	18,2	19,2	23,2	23,6
Urban fabric	1,5	2,4	2,5	2,7

B. LUCC and Water Quality Degradation in Castelo de Bode Dam

Assessing main LUCC in Continental Portugal, lead us to better understand these LULC impact on water quality.

Castelo de Bode Dam reservoir is the most relevant drinking water public supply within the Portuguese context. It is located in the Tagus River Watershed. For this work the drainage area considered is the area that includes the water catchment drainage upstream of the Zêzere river watershed (an area of 396434 ha). Within this area it has been a statistical based model, and established a correspondence between the LUCC and the variation of water quality parameters (WQP), considering the four years that covers the LUCC. The most important LULC of this drainage area in the 1990s were forests, however, this type of LULC decreased since then and the scrub and/or herbaceous vegetation associations became predominant in 2012 (Table 2).

TABLE II. LULC OF DRAINAGE AREA OF CASTELO DE BODE DAM (ZÊZERE WATERSHED) IN DIFFERENT MOMENTS. AREA (%) OBTAINED BY CORINE LAND COVER DATA.

LULC	Year			
	1990	***2000***	***2006***	***2012***
Arable land	4.0	4.0	3.9	4.0
Artificial, non-agricultural vegetated areas	0.01	0.01	0.01	0.01
Forests	41.9	44.0	28.5	31.0
Heterogeneous agricultural areas	15.2	14.9	14.7	13.8
Industrial, commercial and transport units	0.0	0.1	0.2	0.2
Inland waters	1.3	1.3	1.3	1.3
Mine, dump and construction sites	0.1	0.1	0.0	0.1
Open spaces with little or no vegetation	4.7	3.8	2.8	2.4
Pastures	0.1	0.1	0.1	0.3
Permanent crops	1.7	1.7	1.7	2.3
Scrub and/or herbaceous vegetation associations	30.8	29.4	46.0	44.0
Urban fabric	0.4	0.7	0.7	0.7

Analyzing more deeply the available data we can observe an increase in built area during the first decade.

The water quality parameters (WQP) are shown in Table 3.

TABLE III. WATER QUALITY PARAMETERS OF CASTELO DE BODE DAM (ANNUAL AVERAGE).

WQP	Year			
	1990	**2000**	**2006**	**2012**
BDO 5 days (mg/l)	1.43	2.44	3.08	3.00
Total Lead (mg/l)	0.027	0.002	0.004	0.003
Total Coliforms (MPN/100ml)	2839.4	320.7	243.5	75.6
Conductivity in laboratory at 20ºC (µS/cm)	59.0	76.6	85.3	70.8
Color (PtCo)	5	6	9	7
Phenols (mg/l)	0.001	0.008	0.012	0.001
Total Nitrogen (mg/l NO_3)	3.45	1.73	1.84	1.27
Total Nitrite (mg/l NO_2)	0.03	0.02	0.01	0.02
pH - Field	7.46	7.06	7.70	7.24

WQP were obtained through data recorded by automatic stations located in Castelo de Bode dam reservoir (Alb. de Castelo Bode - 16H/03; Cabeça Gorda - 16H/06; Constância - 17G/03; Colmeal - 16H/05) and thought laboratorial analysis, data provided by the National System of Hydrological Decade Resources of Portugal.

The correlation coefficients obtained and presented in Table 4, show a high positive correlation between soil artificialization and the variation of BOD5 (Biochemical Oxygen Demand). These results indicate that the more soil artificialized the greater the biochemical oxygen demand in the waters of the Castelo de Bode Dam reservoir.

On the other hand, variations of lead content, total coliforms, total nitrogen and total nitrites in water shows positive correlation with mine, dump and construction sites and open spaces with little or no vegetation, might be explained by the high use of leaded materials in the construction of infrastructures, entrainment of materials resulting from mining extraction and also by the presence of waste dumps.

Water color changes is positively correlated with variation in areas of industrial, commercial and transport units, as with associations of scrub and/or herbaceous vegetation. These LULC types influence the amount of materials dragged by surface runoff leading to reduction of water quality downstream.

Another relevant conclusion concerns the positive correlation between phenols variation and urban fabric industrial, commercial and transport units, or artificial, non-agricultural, vegetated areas.

The positive correlation between arable land and total nitrites might be explained by previous application of chemical fertilizers in these soils and agrochemicals in these cultures.

All this assumptions demonstrate that LUCC influence water quality and consequently water Framework Directive accomplishment could benefit from this data management insight approach.

C. Interative Model and Results Dessimination

These data models and presented results are being integrated into a WebGIS databases in order to take advantage of Information and Communication Tools (ICT) to assure government action integration and public participation can be effective. Currently the data harmonization (according to the INSPIRE Directive) is being

TABLE IV. CORRELATION COEFFICIENTS FOR THE WQP IN WATER OF CASTELO DE BODE DAM RESERVOIR AND AREAS OF LULC OF DRAINAGE AREA.

LULC	WQP								
	BDO 5 days (mg/l)	***Total Lead (mg/l)***	***Total Coliforms (MPN/100ml)***	***Conductivity in Laboratory at 20ºC (µS/cm)***	***Color (PtCo)***	***Phenols (mg/l)***	***Total Nitrogen (mg/l NO_3)***	***Total Nitrite (mg/l NO_2)***	***pH - Field***
Arable land	-0.62	0.62	0.60	-0.95	-0.85	-0.99	0.43	0.93	-0.35
Artificial, non-agricultural vegetated areas	0.93	-1.00	-1.00	0.84	0.73	0.54	-0.97	-0.88	-0.22
Forests	-0.77	0.42	0.52	-0.53	-0.80	-0.28	0.50	0.46	-0.59
Heterogeneous agricultural areas	-0.73	0.60	0.67	-0.21	-0.35	0.23	0.78	0.23	0.20
Industrial, commercial and transport units	1.00	-0.92	-0.95	0.82	0.85	0.49	-0.93	-0.83	0.05
Inland waters	0.98	-0.97	-0.99	0.87	0.83	0.56	-0.95	-0.88	-0.05
Mine, dump and construction sites	-0.93	0.80	0.83	-0.94	-0.98	-0.74	0.74	0.91	-0.35
Open spaces with little or no vegetation	-0.96	0.80	0.86	-0.64	-0.75	-0.25	0.89	0.63	-0.07
Pastures	0.60	-0.51	-0.57	0.06	0.18	-0.39	-0.71	-0.08	-0.32
Permanent crops	0.51	-0.37	-0.45	-0.07	0.10	-0.50	-0.60	0.05	-0.25
Scrub and/or herbaceous vegetation associations	0.80	-0.46	-0.56	0.54	0.81	0.27	-0.55	-0.48	0.55
Urban fabric	0.95	-0.99	-1.00	0.83	0.75	0.51	-0.98	-0.86	-0.19

done in order to its dissemination using relevant WebGIS platforms (e.g., SmartOpenData, DGT WebSIG - SNIG and IGEO, eEnvPlus, TerAGUA).

These results show the relations between LUCC and water quality changes. The platform being developed can be used as a knowledge base for all decision makers researchers and interested agents involved in water management and territorial planning and survey. The interactivity between producers and users of these data (including the stakeholders) in Web platforms will enable the improvement of information production efficiency, and will lead to increased knowledge in an integrated and harmonized way.

IV. Final Considerations

The results presented in this article show LULC is tightly related to water quality. In fact, LUCC within a watershed, namely those related to anthropogenic activities explain to a large extent the impacts on the quality. This can be verified in the Castelo de Bode Dam reservoir between 1990 and 2012.

In this sense, classifying LULC in terms of their impact in water quality degradation, integrating all assumptions in a WebGIS collaborative portal has high relevance to achieve efficiency in water and land use management domains.

The developed work points to the relevance of sharing knowledge and integrating data in a structured way, that users can address and combine to evaluate land use planning alternatives and their impact on water quality. Data providers can also receive inputs from one another and from relevant users, namely governmental entities with clear attributions in these fields, allowing them to be effective in data collection and integration.

The increase in area of certain types of LULC has an effect on water quality and contributes to the increase on the concentration of pollutants drained or dragged to the main water reservoirs. This study is a highly demanding task that needs a knowledge base complete and up to date. It is essential to extend this approach to other regions so that water management efficiency will increase and the effectiveness of public investments can lead to water stress decline supported in the long term by land use planning efficiency.

All these concerns have to be in line with data collection policy, existing data integration, and well balanced management tools development.

References

[1] A. Erola, and T. Randhir, “Watershed ecosystem modeling of land-use impacts on water quality,” Ecological Modelling, Vol. 270, pp. 54– 63, 2013.

[2] B. Meneses, “The influence of forest fire on water quality of São Domingos stream located in the Western Region of Portugal,” Master Thesys, ISA-UL, Lisbon, 2013.

[3] D. Ahearn, R. Sheibley, R. Dahlgren, M. Anderson, J. Johnson, K. Tate, “Land use and land cover influence on water quality in the last free-flowing river draining the western Sierra Nevada, California,” Journal of Hydrology, Vol. 313, pp. 234–247, 2005.

[4] J. Fiquepron, S. Garcia, and A. Stenger, “Land use impact on water quality: Valuing forest services in terms of the water supply sector,” Journal of Environmental Management, Vol. 126, pp. 113-121, 2013.

[5] K. Charvat, S. Barvika, and M. Alberts, “Linked Open Data for Environmental Protection in Smart Regions – the New Challenge for the Use of Environmental Data and Information,” Proceedings REAL CORP 2014, Tagungsband, Vienna, Austria, pp. 367-376, May 2014.

[6] M. Mendoza, E. Granados, D. Geneletti, D. Pérez-Salicrup, V. Salinas, “Analysing land cover and land use change processes at watershed level: A multitemporal study in the Lake Cuitzeo Watershed, Mexico (1975e2003),” Applied Geography, Vol. 31, pp. 237-250, 2011.

[7] M.J. Vale, “Web based Colaboratory to support water resouces management and landuse Iintegrated planning,” PhD Dissertation, ISEGI, UNL, Portugal, 2002.

[8] M.J. Vale, R. Saraiva, B.M. Meneses, R. Reis, and P. Patrício, “INSPIRE’d land use planning - Dynamic indicators to improve planning achievements,” 8th INSPIRE Conference, Aalborg, Denmark, June 2014.

[9] P. Archer, K. Charvat, M. Cruz, A.L. Alós, J. Estrada, M. Tuchyna, M.J. Vale, and B.M Meneses, “INSPIRE and Linked Open Data for Agro forestry Management – The SmartOpenData Project Approach,” 8th INSPIRE Conference, Aalborg, Denmark, June 2014.

[10] R. Carey, K. Migliaccio, Y. Li, B. Schaffer, G. Kiker, and M. Brown, “Land use disturbance indicators and water quality variability in the Biscayne Bay Watershed, Florida,” Ecological Indicators, Vol. 11, pp. 1093–1104, 2011.

[11] R. Reis, B.M. Meneses, R. Saraiva, M.J. Vale, J. Estrada, and M. Cruz, “Open Linked Data in Portugal - Contribution for INSPIRE Implementation,” 8th INSPIRE Conference, Aalborg, Denmark, June 2014.

[12] R. Saraiva, B.M. Meneses, M.J. Vale, and R. Reis, “COS and land use planning: Open data towards planning efficiency,” Proceedings book of the Environmental Information Systems and Services - Infrastructures and Platforms, ENVIP, Austria, 2013.

[13] V. Rodriguez-Galiano, and M. Chica-Olmo “Land cover change analysis of a Mediterranean area in Spain using different sources of data: Multi-seasonal Landsat images, land surface temperature, digital terrain models and texture,” Applied Geography, Vol. 35, pp. 208-218, 2012.

Implementations of Spatio-Temporal Data Structure for Geographic Information System

Kazuaki Iwamura
Information and Telecommunication Systems Company
Hitachi, Ltd.
Tokyo, Japan
e-mail: kazuaki.iwamura.wx@hitachi.com

***Abstract*—A new spatio-temporal data structure for a geographic information system (GIS) is introduced and employed for various applications. First, the concept for the spatio-temporal data structure is established. The effectiveness of the data structure for management and visualization of real-world states is then demonstrated. To outline its effectiveness, real-time map update and level-of-detail representation are described. The data structure is implemented in a new type of GIS platform, denoted as a four-dimensional GIS (4D-GIS). 4D-GIS applications to social infrastructure management and monitoring are also introduced. 4D-GIS is expected to contribute to the solution of real-world problems.**

Keywords-spatio-temporal; data structure; data integration; four-dimensional GIS.

I. Introduction

Numerous geographical information systems (GISs) have been employed for various applications and used for data management [1]. GIS platforms also have evolved in accordance with developing technology. For example, platforms that can manage 3D data or integrate numerous types of data have made important applications possible. Thus, users are enabled to employ GISs as problem solution tools.

This paper focuses on spatio-temporal and integrative structures to visualize and solve real-world problems. Actually, many real-world problems are concerned with two-dimensional (2D) and three-dimensional (3D) data that include temporally changing elements and aspects, which are recorded in various types of documents such as maps and diagrams. For this purpose, common data structures and integration methods are introduced to solve various problems.

Spatio-temporal data modeling has hitherto been proposed and applied to various fields [2]-[4]. In [2], a three-domain model comprising semantic, temporal, and spatial domains is proposed to realize complex query compositions. In [3], two temporal dimension models are introduced to represent actual and planned events. In [4], a data structure for spatio-temporal queries is proposed. In this paper, a new type of spatio-temporal data structure that can represent 2D, 3D and their time changes and their applications is introduced.

In Section II of this paper, the basic formulation of the proposed spatio-temporal data structure is established. In Section III, real-time map updates and level-of-detail (LOD) representation, which are data visualization control methods, are described to illustrate some spatio-temporal aspects. In Sections IV, V, and VI, the structure is implemented in a new type of GIS platform, denoted as four-dimensional GIS (4D-GIS), and applications to social infrastructure management and monitoring are introduced. In Section VII, the effectiveness of the proposed spatio-temporal data structure is discussed and a conclusion of the work follows.

II. The Spatio-Temporal Data Structure

In this section, the basic formulation of the proposed spatio-temporal data structure and its extensions to 3D and temporal changes are established.

A. The Structure of Spatio-temporal Shapes

Spatio-temporal shapes are coordinate data whose dimensions are 2D and 3D in addition to temporally correlated changes. An illustrative structure is shown in Fig. 1.

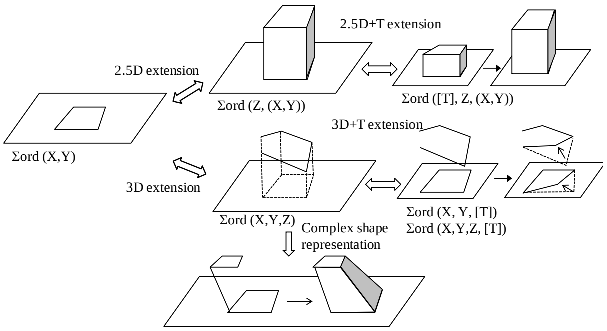


Figure 1. Spatio-temporal data structure

The basic shape of the structure is 2D. The 2D structure is written as follows in (1).

$$\Sigma_{\mathrm{ord}}\,(X, Y) \quad . \tag{1}$$

Here, Σ_{ord} does not imply a summation of numerical values but indicates ordered (X, Y) coordinate sequences. All shape data whose dimensions are greater than 2D are obtained by the expansion of this 2D structure. 3D shapes are extended by inserting a new dimension for height or depth. This extension takes one of the two forms given in (2) and (3).

$$\Sigma_{ord}\,(Z, (X, Y)) \quad , \tag{2}$$
$$\Sigma_{ord}\,(X, Y, Z) \quad . \tag{3}$$

Structure (2) represents a 2.5-dimensional (2.5D) structure, where Z is a common height/depth dimension for all (X, Y) coordinates. Using (2), 3D structures such as buildings can be represented. On the other hand, for structure (3), all (X, Y) coordinates include individual Z coordinates. For example, sewer pipes, which are inclined, would be represented by 3D polylines using structure (3).

Spatio-temporal structures are expansions of structures (1), (2), and (3). For structures (1), (2), and (3), time data are attached as follows in (4), (5), and (6), respectively.

$$\Sigma_{ord}\,(X, Y, [T]) \quad , \tag{4}$$
$$\Sigma_{ord}\,([T], Z, (X, Y)) \quad , \tag{5}$$
$$\Sigma_{ord}\,(X, Y, Z, [T]) \quad . \tag{6}$$

Here, [T] is a time value that is attached only to changing coordinates. Because time is a continuous value, the time attributes are attached as follows.

- creation start time and creation end time,
- elimination start time and elimination end time.

A given start time and end time are paired values.

The 4D-GIS serves as a spatio-temporal data integration platform, and all data structures (2)–(6) can be expanded from the base structure (1) according to dimensions required by a given application. The merits for adopting the above spatio-temporal data structures are as follows.

- Because past data is not deleted, time change sequence data remains available for applications.
- Data with future time change sequence data can be stored for planning applications.
- Because changed structures remain in a coordinate sequence, it is easy to locate the point in time when changes occur.
- Because only coordinate difference data is stored, data storage does not become large.

B. Extension to Complicated Shapes

Real-world object shapes tend to be more complex than those presented above. Shapes that are more complicated can be represented by a combination of structures given by (1)–(6). The equation for combination is as follows in (7).

$$\text{Shape} = F_1\,(S_1\,(t), S_2\,(t), L_1(t)) + F_2\,(S_2\,(t), S_3\,(t), L_2(t)) + \ldots. + F_n\,(S_n\,(t), S_{n+1}\,(t), L_n\,(t)) \quad . \tag{7}$$

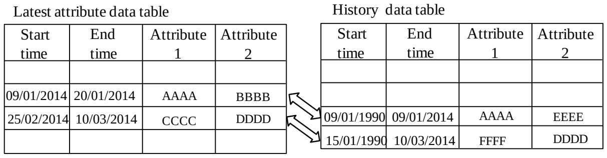
Here, S_i is a control surface, L_i is a control line, and t is a parameter. S_i and L_i have the same structures given in (1)–(6), and time varying representations are also possible. Full shape data is created by interpolating between control surfaces S_i and S_{i+1} along a control line L_i. This structure can be denoted as homotopical data. The word homotopy is a mathematical term in topology describing smooth deformation/connection between two shapes.

Two types of homotopical data generation exist.

- Interpolation of a control line by curve fitting, where control surfaces are attached to both ends of a control line and a fit curve is generated among a control line. An example is shown later in Fig. 6.
- Interpolation by model shapes along a control line, where model shapes are generated and connected along a control line successively. An example is shown later in Fig. 3.

C. Spatio-Temporal Attribute Data

Attribute data, which are connected by a position or by map objects, are also managed by the temporal structure. Fig. 2 shows the table structure of attribute data in a relational database.



Latest attribute data table

Start time	End time	Attribute 1	Attribute 2
09/01/2014	20/01/2014	AAAA	BBBB
25/02/2014	10/03/2014	CCCC	DDDD

History data table

Start time	End time	Attribute 1	Attribute 2
09/01/1990	09/01/2014	AAAA	EEEE
15/01/1990	10/03/2014	FFFF	DDDD

Figure 2. Attribute data and their history management

This structure comprises two table types. One is the latest attribute data table and the other is the history data table. The latest attribute data table stores the latest data and the history data table stores records to be changed.

Each record in both data tables includes “start time” and “end time” terms. When the attributes in the record fields in the latest attribute data table are to be changed, data in these records are moved to the history data table and the available periods are attached according to the start time and end time. Then, fields of the latest attribute data table are changed by the newly obtained data. When GIS users want to refer to past attribute data, the corresponding records in the latest data table and the history data table are exchanged. Thus, histories of attribute data are managed and used.

D. Spatio-temporal Linear Feature Data

Linear features are concerned with elongated configurations such as roads and trunk gas/oil pipelines. Data for these facilities in GISs are managed by a distance parameter [5]. Moreover, linear features that change according to time can be denoted as spatio-temporal linear features. For example, defect positions such as pipeline corrosion are managed using distances from a reference point, and corrosion expands gradually with time.

Using the complicated shape representation given by (7), shape data for large-scale social infrastructures can be generated. An example of road shape generation is shown in Fig. 3.

The shape generation steps are as follows.

Step 1: facility parameters are retrieved according to the distance value obtained from the attribute database.

Step 2: templates given as model shapes are generated by referring to facility parameters.

Step 3: model shapes are aligned along a control line and connected. Then, the full shapes are generated.

Merits of this process of shape generation are as follows.

- The linear feature model is generated using attributes, and is therefore a compressed type of data.

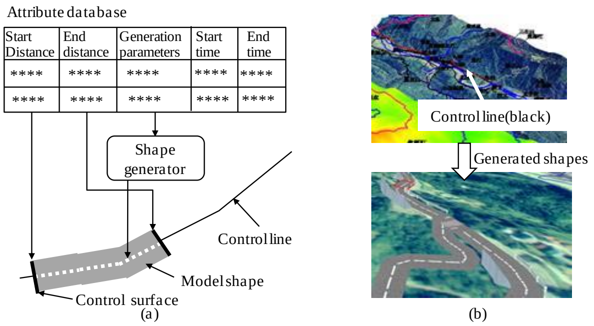


Figure 3. Road shape generation: (a) Principle, (b) Example

- When facilities are changed, it is sufficient to update only the parameters of the facility attribute data because new shapes can be generated using the updated parameters.

As an example, when the method is applied to updating map data for a sound-proof wall of an elevated causeway, the update cost using linear features is about 1/3 that when using full shape data.

III. Real-time Map Update and level-of-detail Control

In this section, two remarkable uses of the proposed spatio-temporal data structure and the homotopical data structure are presented. One use is in real-time map update and the other is LOD control in shape visualization.

A. *Real-time Map Update*

Map data becomes antiquated immediately after creation because the real world changes rapidly. Therefore, timely updating map data in accordance with real-world changes is preferable. On the other hand, the obtained map data should be authorized by the responsible authorities, and updates of original maps can require considerable time for completion. TABLE I lists the update cycle times of gas and water supply facility maps prepared by a local government of Japan.

TABLE I. Update Cycles of Facility schematics

Drawings/Maps	Scale	Update cycle	Time to update
Gas supply facilities	1/500	3 times/Year	2–5 months
Water supply facilities	1/500	Once/Year	2–5 months
Water facilities' master	1/500	Once/Year	One year
CAD Drawings	1/200	Once/Year	One month

Because infrastructure such as gas pipelines are repaired or exchanged daily, map data in which the latest infrastructural configurations are reflected should be created in a timely manner. Thus, a new method for the real-time map update was developed. The developed system is illustrated in Fig. 4, and an example of map update is shown in Fig. 5.

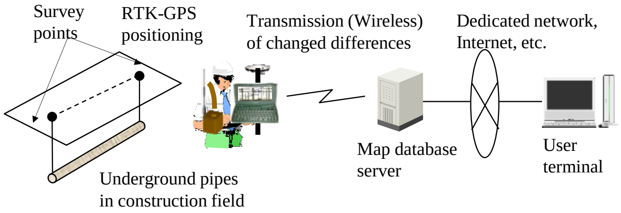
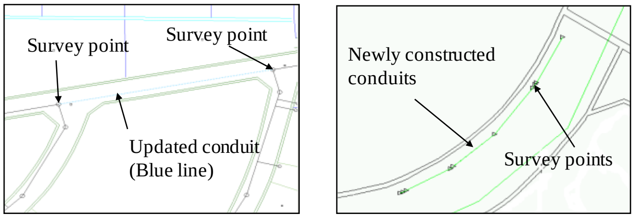

Survey points
RTK-GPS positioning
Transmission (Wireless) of changed differences
Dedicated network, Internet, etc.
Underground pipes in construction field
Map database server
User terminal

Figure 4. System for real-time map update



Partial update of conduit   Newly constructed conduit

Figure 5. Example of real-time map update

As shown in Fig. 4, real-time kinematic global positioning systems (RTK-GPS) are employed to obtain facility position coordinates such as for pipes, which represent a segment of infrastructure. Then, the configuration coordinates of this segment with time values are transferred to a map database server, and the map data is immediately updated according to the spatio-temporal schema.

At this point, a comment is required. The obtained coordinate data are embedded into the map data. However, the updated maps are not official until authorized. Thus, the updated maps are temporary. To accommodate authorization requirements in the spatio-temporal data format, an authorization tag bit is incorporated. Prior to authorization, this tag bit is switched off. After authorization, the tag bit is switched on, allowing 4D-GIS users to discriminate the status of map elements. This method has important merits whereby current real-world states can be evaluated and 4D-GIS users can employ the latest maps at any time.

B. *Level-of-detail Control*

The LOD establishes design criteria for map visualization control. In this section, two types of LOD design methods that are concerned with accuracy controls are described. One is a 2D case and the other is a 3D case. The LOD control is necessary to visualize maps of different accuracies successively.

1) *Level-of-detail control in 2D visualization*

In this section, the homotopy concept of (7) is applied to curved polyline accuracy controls. When map data is magnified, curved polylines such as contour lines become coarse. However, the resolution of curved polylines should remain consistent even when magnified. One solution to avoid this situation is to prepare a plurality of curved polylines with different levels of accuracy and to select them according to the magnification employed. However, when data are updated, all related data should be updated. Thus, a curve fitting method based on homotopical data interpolation was developed. The fitting equation based on (7) is as follows in (8).

$$\mathrm{Curve}_i(t) = (1 - t)C_i + tC_{i+1} \quad . \qquad (8)$$

Here, Curve_i is a fitting curve for the line L_i, where $0 \leq t \leq 1.$, and C_i and C_{i+1} are circle arcs. The method is also illustrated in Fig. 6 (a).

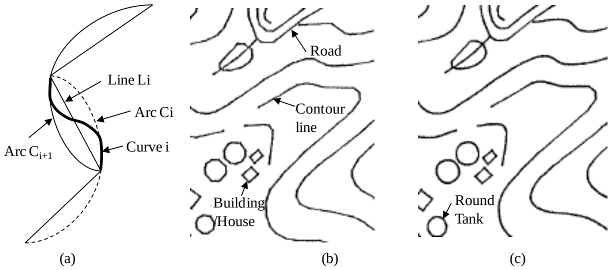


Figure 6. Homotopical LOD control of 2D shapes: (a) fitting principle, (b) shapes before fitting, (c) shapes after fitting.

Polylines might include groups of segments that represent large curvatures. At these segments, curve fitting without controls would result in an unexpected overlap with other shapes. Thus, the four controls listed in TABLE II are imposed.

TABLE II. LOD CONTROL OF 2D SHAPES (CURVE FITTING)

Parameter	**Condition**	**Control**
Line length L_i	$TL_{min} \leq \lvert Li \rvert$ $TL_{max} \geq \lvert Li \rvert$	$C_i \rightarrow L_i$ in (8)
Angles θ of two lines L_i, L_{i+1}	$TA_{min} \leq \lvert \theta \rvert$ $TA_{max} \geq \lvert \theta \rvert$	$C_i \rightarrow L_i$ in (8)
Maximum deviation d between C_i and L_i	$d \leq T_d$	Fitting (8) stops
Ratio r/d (r is radius of C_i)	$r/d \geq T_r$	Fitting (8) stops

Here, T_{Lmin}, T_{Lmax}, T_{Amin}, T_{Amax}, T_d, and T_r are thresholds. T_r is determined uniquely. The consideration wherein squares are not fitted, and therefore, not changed to a circle requires that

$$d = r - \frac{r}{\sqrt{2}} \text{ and } T_d = \frac{r}{d} = \frac{\sqrt{2}}{\sqrt{2}-1} \cong 3.4. \quad . \qquad (9)$$

The result is shown in Figs. 6 (b) and (c).

2) Level-of-detail control in 3D visualization

In 3D urban area visualization, 3D shapes far from a given viewpoint are small and details are degenerated. Thus, display of all 3D data including details is needlessly time consuming, making it is necessary to control the visualization accuracy. For example, shapes near a viewpoint are visualized in detail, whereas shapes far from a viewpoint are eliminated or the details are omitted. Using homotopical data structure, LOD controls can be realized effectively. The design criteria of LOD are listed in TABLE III. Here, T_{L1}, T_{L2}, T_{L3}, T_H, and T_S are thresholds.

TABLE III. LEVEL-OF-DETAIL CONTROL OF 3D SHAPES (OBJECTS)

Parameter	Condition	Result
Distance L from a viewpoint to a shape	$L \geq T_{L1}$	Objects whose heights are lower than T_H are deleted.
	$T_{L2} \geq L \geq T_{L3}$ ($T_{L2} \geq T_{L3}$)	Small lines whose lengths are less than T_s are deleted. When an object comprises a group of objects, the group is united into a single average object.

Fig. 7 illustrates the application of these control criteria.

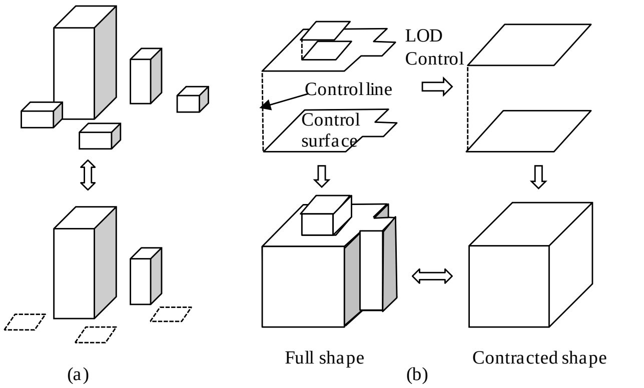


Figure 7. Homotopical level-of-detail (LOD) control of 3D shapes: (a) object control and (b) deformation.

Fig. 7 (a) presents an example of object control, where objects far from a viewpoint are eliminated. Fig. 7 (b) is an example of deformation, where detailed segments of objects are removed because the details fade as objects recede from a viewpoint.

The merits of the homotopical method in LOD controls are as follows.

- The amount of data required to construct models are significantly reduced. For example, using this method to construct a 3D model of the crowded city-center of Tokyo reduces the required data size by 28% on average relative to the use of full shapes.
- By managing only the original homotopical data, any LOD data can be generated by deformation of control surfaces.
- Temporal changes are also represented by attaching time values in control surfaces and control lines.
- Smooth viewpoint changes are also realized.

IV. TRUNK GAS PIPELINE INTEGRITY MONITORING

In Sections IV, V, and VI, applications using spatio-temporal data to trunk gas pipeline integrity monitoring, communication facility management, and power consumption monitoring are introduced. In this section, trunk gas pipeline integrity monitoring is presented.

A. Pipeline Integrity Data

For trunk gas pipelines, transport pressures are greater than 4 MPa. Numerous trunk pipelines whose lengths are greater than 1000 km pass through several countries. When pipelines are buried underground and not adequately maintained, defects such as corrosion appear, and the wall thicknesses of pipes decreases over time. These conditions induce pipeline collapse. Thus, GISs are introduced for maintenance support to protect against pipeline collapse and possible explosions. The requirements for GISs are as follows [6].

- Representation of overall pipeline conditions. Data reflecting the overall pipeline status should be integrated and retrieved by map interfaces.
- Prediction of potential risks. Potential risks such as deterioration of pipes caused by expanding defects should be readily detected.
- Contribution to stable and safe gas transport. The stress concentrations on defects should be calculated.

In TABLE IV, pipeline integrity data necessary to meet the above requirements for pipeline monitoring are listed.

TABLE IV. DATA FOR PIPELINE INTEGRITY MONITORING

Name of data	Contents
Maps	Pipeline configurations based on 3D coordinates
Facility construction records	Construction specifications such as pipe section length, wall thickness, and pipe diameter
Defect inspection data	Corrosion dimension and positions obtained by the autonomous in-line inspection robot called a "pig"
Protection data	Cathode voltage
Gas transport parameters	Transport pressures at discharges and suctions at compressor stations. Parameters can be obtained from a supervisory control and data acquisition (SCADA) system
Survey results	Soil conditions, facility conditions obtained by digging

Pipelines are often laid on undulating terrain. Thus, pipeline configuration data must possess 3D features. Nearly all data are managed by distance parameters. Moreover, according to a given lapse of time, pipelines located on soft soils tend to sink and defects tend to expand in the absence of repair. Thus, nearly all integrity data are managed as spatio-temporal linear features.

B. Safety Diagnoses

Corrosion expansion is concerned with not only soil conditions, but also pipeline materials and transport pressures. The number of areas of corrosion in/on pipelines is occasionally greater than 1,000 per 1 km in the absence of maintenance. However, not all incidents of corrosion are dangerous. Estimating the risks associated with corrosion is important. For this purpose, a safety diagnosis function was incorporated with the 4D-GIS.

The corrosion expansion speed is obtained by extracting size differences between two or more corrosion history data points. In the absence of corrosion history data, the corrosion expansion speed V can be determined by (12) as follows.

$$V = V_0\,(1+E\varepsilon_0)\,\exp\,\{f\,(T,\,\varepsilon_1,\,M)\} \quad . \qquad (12)$$

Here, V_0 is the corrosion speed under no stress conditions, E is an environmental parameter, $\varepsilon_0/\varepsilon_1$ is the volumetric/average strain, f is the stress function, T is the fluid temperature, and M is a pipeline material parameter. By comparison with real data, corrosion expansion is found to coincide with measured values within an error to the order of 0.01 mm.

In Fig. 8, a screen image of the 4D-GIS safety diagnosis function is shown.

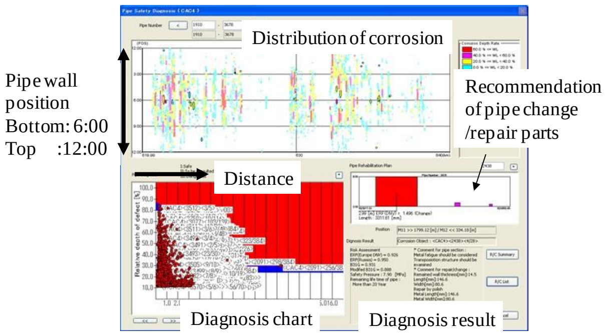


Figure 8. Safety diagnosis

The upper graph in the figure is the corrosion distribution diagram. The bottom left is the corrosion diagnosis chart. The bottom right is the update/repair recommendation and the diagnosis result. In the safety diagnosis, all instances of corrosion in/on the selected pipe sections are projected after the prediction of the current states using past corrosion data or (12). When corrosion is in the red area, exchange or repair of the corroded pipe section is recommended. The blue area represents the warning zone and the white area is the safety zone. Corrosion in the white zone is safe. Moreover, the remaining life time of the selected pipe sections are also predicted using (12).

C. Analysis of Stress Concentration

When the wall thickness of a pipeline section thins owing to severe corrosion, stresses caused by transport pressures can induce pipeline failure and potential explosions. Therefore, monitoring of the transport pressure distribution and stress concentrations are highly recommended. An extraction of stress concentration on pipeline sections is shown in Fig. 9.

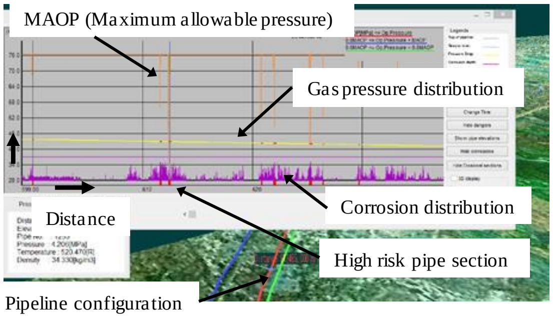


Figure 9. Calculation of pipeline stress concentration

In Fig. 9, red bands means high risk pipe sections. Orange graph means maximum allowable operation pressures (MAOP). Stress concentrations depend on the transport pressure, corrosion size, and pipe dimension. It is convenient to transform stress values to allowable pressure values. Stress values are calculated by the stress-strain equation. An allowable pressure P_{max} is calculated using the following equation (13).

$$P_{max} = 4\,\sigma\,\delta/D \quad . \qquad (13)$$

Hear, σ is the axial stress, δ is the modified wall thickness, and D is the modified pipe diameter. When P_{max} /MAOP > α, pipe sections with these Pmax values have high risks. α = 0.5 for safety.

When a bypass pipeline, denoted as looping, is utilized, the transport pressure is reduced by an operational method that divides the gas flow into a main pipeline and a looping pipeline.

V. Optical Communication Facility Management

In this section, an optical communication facility management is introduced. Some local governments in Japan have constructed optical communication networks in sewer underdrains and pipes for data communication among sewage disposal plants. For maintaining communication facilities, engineers refer to numerous types of documents. A listing of these reference documents is given in TABLE V.

TABLE V. Documents for communication Facility Maintenance

Type of document	Content
Map	Communication routes and their locations
Facility diagram	Connections of network equipment
Network diagram	Schematic communication routes
Fiber connection diagram	Routes of fibers cores included in cables
Facility data	Specifications of facilities

Maps, diagrams, and facility data are altered according to network expansion. Moreover, future networks are also input. Thus, spatio-temporal data management was adopted.

In maps, geographical locations of communication routes are presented. However, equipment such as fiber connection boxes is not represented. On the other hand, other diagrams are topological, and connection sequences of network equipment are given. When documents are used, the following problems are encountered.

- When maps and diagrams are separately referenced, facility sequence information cannot be obtained from maps and facility locations cannot be obtained from diagrams.
- Fiber connection diagrams describe the connection routes of fibers. Because the number of fibers in a cable is large, drawing diagrams manually is difficult.

Facility managers and engineers use all types of documents in daily maintenance work, and the 4D-GIS can solve the problems described above by integrating the different types of documents employed.

A. *Categorical Connection among Documents*

The method employed for the integration of maps and diagrams is described. Both types of documents are connected according to category theory [7][8]. Category is a mathematical concept that describes the relations of homomorphic structures.

The connection method among two structures is illustrated in Fig. 10 (a). In the figure, “Map” indicates map objects, “Dgm1” and “Dgm2” indicate two different diagram objects, and “Attr” indicates attributes. Relations among maps and diagrams are illustrated in Fig. 10 (b).

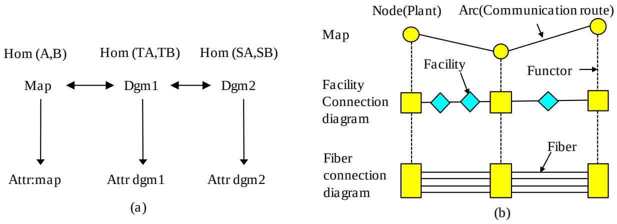


Figure 10. Connection between documents based on categories: (a) connection structure, (b) interconnections among documents.

Nodes represent plants and links between two communication routes. The nodes in maps and diagrams are connected beforehand because the number of nodes is large (about 100). Routes between nodes can be calculated by a path-finding algorithm.

The mappings between nodes and between links can be denoted as functors. Functors are bridges between two structures. A functor τ is a projection of nodes and links as follows in (14).

$$\tau\colon \mathrm{Hom}(A,B) \rightarrow \mathrm{Hom}(TA,TB) \qquad (14)$$

Here, Hom(A,B) is a projection and corresponds to a route from node A to node B, whereas TA and TB are mapped nodes in the other document that correspond to node A and node B, respectively. In a facility connection diagram, facility connection sequences are described. However, map users cannot retrieve the sequences of facilities and their specifications because these data are not attached to maps but to diagrams. Thus, functors play important roles as connectors among documents.

The outline of steps for retrieving communication routes and facility data by linking between a map and diagram are described below.

Step 1: selection of two nodes on a map.
Step 2: finding corresponding nodes on a diagram by a functor.
Step 3: finding the related route on a diagram, as calculated using a path-finding algorithm.
Step 4: retrieval of facility data on the selected route from the database.

An example of related route finding based on the above steps is shown in Fig. 11.

Following the finding of the route on the map (the red color route in the map), the corresponding route in the facility diagram (the red color route in the facility diagram) is detected.

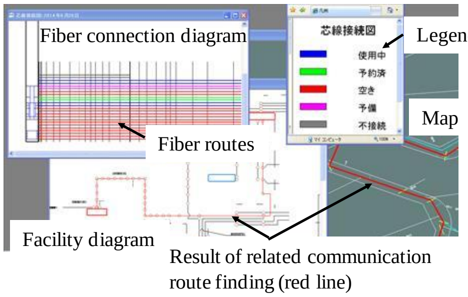


Figure 11. Related route finding and generation of fiber connections diagram

B. *Automatic Fiber Route Diagram Generation and Implicit Data Management*

Presently, fiber connection diagrams are drawn manually. However, drawing fiber connection diagrams is difficult because a great many fiber core lines, occasionally more than 1000, are included in a single cable. Using the functor approach, fiber connection diagrams can be generated automatically. Moreover, a new database method, referred to as implicit data management, is shown to contribute to the retrieval of fiber connection data easily.

Two fiber cores are connected by melting or with connector boxes, and a very great amount of connection data must be managed effectively. However, to input connection data manually is time consuming. Every fiber core has an index such as 1-2, 2-3. The former number represents the cable case number, whereas the latter number represents the fiber core number. Two fiber cores with the same index are ordinarily connected whereas few fibers with different indices are connected exceptionally. Thus, only connection data with different indices is stored in the database and ordinary connection data is not stored. In this manner, when an index pair does not exist in the database, fibers with the same indices are connected. This data registry method is herein denoted as implicit data management. The total data size is thereby reduced by greater than 1/5. A result of fiber connection diagram generation is shown in Fig. 11. The fiber connection diagram that corresponds with the finding route (the red color route in the map) is generated.

VI. Big Data Processing in Power Monitoring

Recently, numerous smart city projects have been planned and executed. In smart cities, effective monitoring of power supply and consumption is pursued. For this purpose, energy consumption data is collected by equipment such as smart meters and analyzed to determine consumption trends. The data occasionally comprise greater than 10 million individual elements. It is a technical challenge to develop methods to handle such large power consumption data sets, which can be denoted as big data. For this purpose, the 4D-GIS adopted in-memory techniques. In-memory techniques employed in 4D-GIS are shown in Fig. 12 and described as follows.

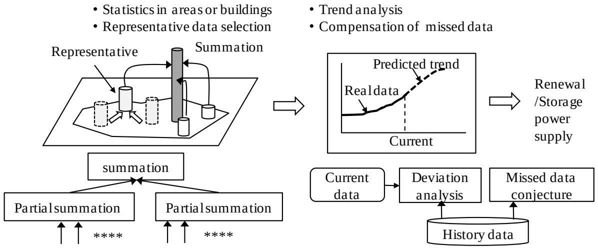
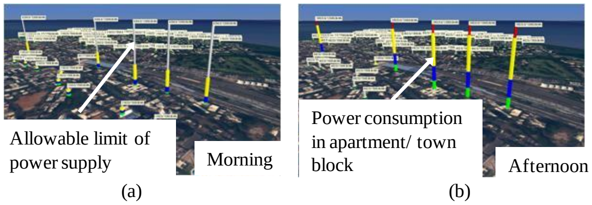


Figure 12. In-memory processing techniques employed in 4D-GIS

- Capture of statistics over areas or for individual buildings obtained from the summation of power consumption values in designated areas or for building objects in maps.
- Representative data selection from, for example, houses that show similar consumption values and patterns. In this example, a representative house is selected and the consumption data of other houses are omitted. Thus, the amount of collected data is greatly reduced.
- Trend analyses using historical data from which future prediction is executed.
- Compensation for missing data. All power consumption data obtained in a specified time interval is managed within the database by spatio-temporal schema. However, when communication facilities fail, some data becomes unavailable, and values of missing data can be extrapolated from the consumption history stored in the database.

The visualization reflecting changing conditions of electricity consumption is shown in Fig. 13.


Allowable limit of power supply
Morning
(a)
Power consumption in apartment/ town block
Afternoon
(b)

High Resolution Satellite Image:ⒸDigital Globe and Hitachi Solutions, Ltd.
The screen images were created using flood simulator by Hitachi Power Solutions Co., Ltd.

Figure 13. Visualization of power consumption monitoring: (a) morning, (b) afternoon.

In the figure, green, blue, and yellow segments of the columns represent energy consumption by different equipment in apartments or town blocks. The gray segments of the columns represent the allowable limits of power consumption. In Figure 13 (a), electricity consumption in the early morning is shown and is low because communities are not particularly active. In Figure 13 (b), electricity consumption in the afternoon is shown, where some of the formerly gray columns have changed to red owing to power consumption close to supply limits.

VII. Discussion and Conclusion

The spatio-temporal structure is a reflection of real-world states. Thus, the 4D-GIS in which this structure is implemented becomes a platform to solve real-world problems. The benefits for the 4D-GIS utilization are as follows.

- 3D and temporal changes are represented by a natural extension of the 2D data structure. Thus, 4D-GIS can be readily extended to other application fields.
- By combination with homotopical representation, complicated data can be constructed effectively. Data size is reduced and the visualization effect denoted as LOD can be implemented.

Hitherto, other types of spatial data structure have been authorized as international standards. Data structures such as web map service (WMS) and web feature service (WFS) are notable examples. These formats are used for data exchanges among GISs. The proposed spatio-temporal structure is native to the 4D-GIS platform. The 4D-GIS imports standard format data such as WFS, converts it into the native format, performs operations on this data, and the results can be exported to other GISs after converting into WFS or WMS data structures.

Numerous future challenges are also evident. In the current implementation, the data structure is applied in the representation of only man-made objects. For further applications, representations of soil and atmosphere are also necessary.

In this paper, a new type of spatio-temporal data structure has been introduced and has been implemented using the 4D-GIS integration platform for various applications. The spatio-temporal data structure has demonstrated the following remarkable characteristics.

- Real-world states can be effectively represented, allowing for realistic solutions to real-world problems.
- Data representation is flexible and many applications unavailable to traditional 2D map base GISs can be realized.

Three applications, trunk gas pipeline integrity monitoring, optical communication facility management, and power consumption monitoring, have demonstrated the effectiveness of the proposed spatio-temporal structure.

ACKNOWLEDGMENT

The author would like to thank Enago (www.enago.com) for the English language review.

REFERENCES

[1] R. W. Greene, “Open access: GIS in e-government”, Esri Press, 2001.

[2] G. Langran and N. R. Chrisman, “A framework for Temporal Geographic Information”, Cartographica, Vo. 25, No.3, pp.1-14, 1988.

[3] M. Yuan, “Use of three-domain representation to enhance GIS support for complex spatiotemporal queries,” Transaction in GIS, Vol 3, No.2, pp.137–159. 1999.

[4] M. F. Worboys, “A unified model for spatial and temporal information,” The Computer Journal, Vol. 37, No.1 pp.26–34, 1994.

[5] K. Iwamura, K. Muro, N. Ishimaru, and M. Fukushima, “4D-GIS (4 dimenional GIS) as spatial-temporal data mining platform and its application to large-scale infrastructure,” 1st IEEE Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM2011), pp.38–43, 2011.

[6] K. Iwamura, A. Mochiduki, Y. Kakumoto, S. Takahashi, and E. Toyama, “Development of spatial-temporal pipeline integrity and risk management system based on 4 Dimensional GIS(4D-GIS)”, Proc. of International Pipeline Conference 2008 (IPC 2008), pp.291–298, 2008.

[7] S. Maclane, “Categories for the mathematician,” Springer-Verlag, 2nd ed. 1997.

[8] J. Herring, M.J. Egenhofer, and A.U. Frank, “Using category theory to model GIS applications”, 4th International Symposium on Spatial Data Handling, pp.820–829, 1990.

ISBN: 978-1-61208-383-4

A Study on the Insurance Premium Rate Map Considering the Natural Disaster Risks in Korea

Jung Ok Kim, Junseok Lee, Jihun Kang
Spatial Information Research Institute
Korea Cadastral Survey Corporation
Seoul, Korea
e-mail: jungok@lx.or.kr, jaslee@lx.or.kr, kangdaejang@lx.or.kr

Abstract—**Recently, the number of natural disasters has increased due to abnormal changes in the weather in Korea. A storm and flood insurance system is available to prevent fallout from these natural disasters. The national storm and flood insurance premium rates are very low and the risk of adverse selection exists because the system is based exclusively on those who live in high-risk areas. To solve these problems, a storm and flood insurance premium rate map is required. In this paper, we investigate domestic and foreign storm and flood insurance maps and extract the common elements of these storm and flood insurance maps. We also created a prototype storm and flood insurance premium rate map of Ulsan in Korea.**

Keywords-natural disaster; hazard risk; insurance map; GIS; insurance premium rate.

I. INTRODUCTION

Recent widespread flooding and other extreme weather events have caused devastating losses across Korea. These losses have been borne by individuals and businesses, local governments, community organizations, the by Korean government. For this reason, the Korean government has developed a storm and flood insurance system which pays benefits in the event of this type of damage [1]. This system was introduced in 2006 to protect businesses and residences from the effects of natural disasters such as typhoons, floods, torrential rain, gales, heavy seas, tidal waves, heavy snowfalls and earthquakes, as well as tsunamis, as shown in Fig. 1 [2]. The storm and flood insurance system is managed by the National Emergency Management Agency in Korea and is administered by private insurance companies. The central (59.5%) and local (25.5%) governments offer partial subsidies for the insurance premiums of customers (15%) to make them more capable of coping with unexpected storms and floods [3].

Generally, standard insurance premium rates are rationally graded or applied according to the magnitude of the risk (the degree of risk). However, present storm and flood insurance in Korea applies the same insurance premium rate regardless of the risk or district, using only one risk grade all 230 cities covered [4]. Therefore, the current insurance premium rate is unfairly discriminatory, as it is supposed to be graded and applied in the standard manner of insurance premium rates based on the level of risk [5]. For example, as shown in Fig. 2(a), the same insurance premium rate is applied despite the fact that zone A has a lower degree of risk than zone B. This study provides a framework for applying different insurance rates depending on the risk of the subject-matter insured, as shown in Fig. 2(b) [4].

This paper aims to present spatial data and analysis methods to assist the creation process of a storm and flood insurance premium rate map. A concurrent aim is to produce a pilot map. The rest of this paper is organized as follows: Section II gives a definition of an insurance premium rate map and demonstrates how to produce such a map for Ulsan in Korea. Section III concludes the paper and summarizes the benefits stemming from its results.

II. INSURANCE PREMIUM RATE MAP

The storm and flood insurance map is the digital map constructed by a computer system as the thematic map that represents the risk of damage from storm, flood and snow.

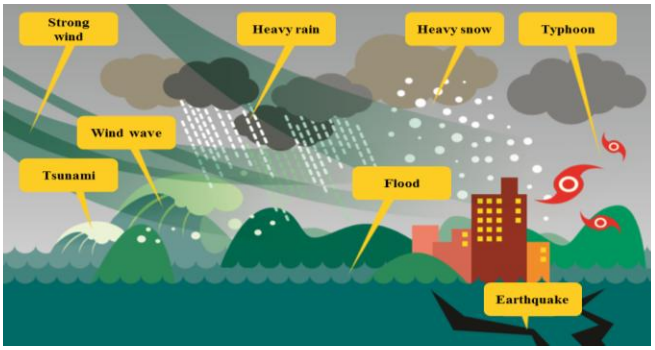


Figure 1. Targets of storm and flood insurance.

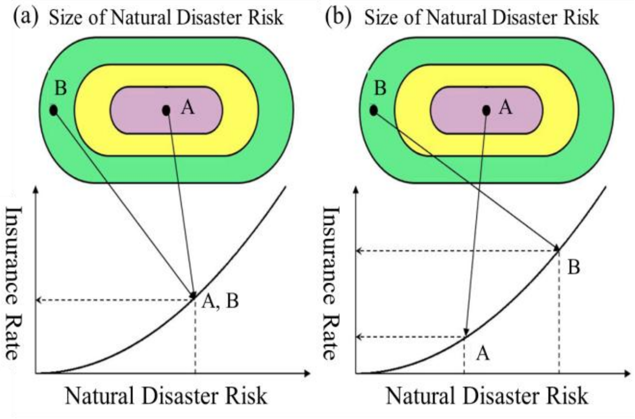


Figure 2. Insurance premium rate system: (a) current (b) revised.

Fig. 3 shows the calculation procedures to determine the storm and flood insurance premium rates. Also shown are the components of the insurance map. The risk levels of three disasters (wind, snow, and water) can be calculated through a grid operation. In this study, we used a grid resolution of 10m, in accordance with the resolution of the DEM (digital elevation model). This work was performed with ArcGIS 10.1 [6].

The completed insurance premium rate map is shown in Fig. 4(a), which shows the final map classified into four grades – safe, alert, dangerous, and very dangerous – in contrast to the existing single premium rate for each of the five districts of Ulsan City.

This map was converted into a KMZ file for the sharing of the data. Then, as shown in Fig. 5, insurance premium rates on Google Earth could be obtained. The data analysis results from the insurance premium rate are also shown above the image of the map.

III. CONCLUSION

We proposed the storm and flood insurance premium rate map. Also we made the prototype storm and flood insurance rate map of the Ulsan Korea. In this paper, we were focused that GIS-based spatial database was constructed for risk zone information about storm and flood damage. It is possible to analyze various disasters and to support disaster prediction for the future.

ACKNOWLEDGMENT

This research was supported by a grant ‘Development of Insurance Rate Map based on National Disaster Risk’ [NEMA-NH-2013-62] from the Natural Hazard Mitigation Research Group, National Emergency Management Agency of Korea.

REFERENCES

[1] J. Lee, I. Lee, J. Kang, and S. Yeon, “GIS technique for analysis of storm and flood insurance premium Rates,” 2014 Conference on Korean Society of Hazard Mitigation, KOSHAM, Feb. 2014, pp. 228–228. Available from: http://db.koreascholar.com/article.aspx?code=268336 [accessed: Dec. 2014].

[2] Korean Government, Stom and flood insurance act, Available from: http://elaw.klri.re.kr/kor_service/lawView.do?hseq=26560&lang=EN [accessed: Dec. 2014].

[3] National Disaster Information Center. Definition of storm and flood insturance. Available from: http://www.safekorea.go.kr/dmtd/contents/sdiw/info/SdiwInfo01.jsp?q_menuid=M_NST_SVC_08_01_01_01 [accessed: Dec. 2014].

[4] H. Lee, S. Park, C. Lee, and Y. Kim, “The point at issue and improvement of natural disaster insurance rate system,” Journal of Korean Society of Hazard Mitigation, Korean Society of Hazard Mitigation, vol. 14 no. 1, pp. 223-231, Feb. 2014, doi:10.9798/KOSHAM.2014.14.1.223.

[5] K. Kim, “A theoretical study on storm and flood insurance in Korea,” KDI Journal of Economic Policy, vol. 33, no. 4, pp. 119-141, Dec. 2011. Available from: http://journal.kdi.re.kr/paper/paper_view.jsp?art_no=2106&pub_no=12230 [accessed: Dec. 2014].

[6] ArcGIS Help 10.1: An overview of the Local toolset, Available from: http://resources.arcgis.com/en/help/main/10.1/index.html#/An_overview_of_the_Local_tools/009z0000007p000000/ [accessed: Dec. 2014].

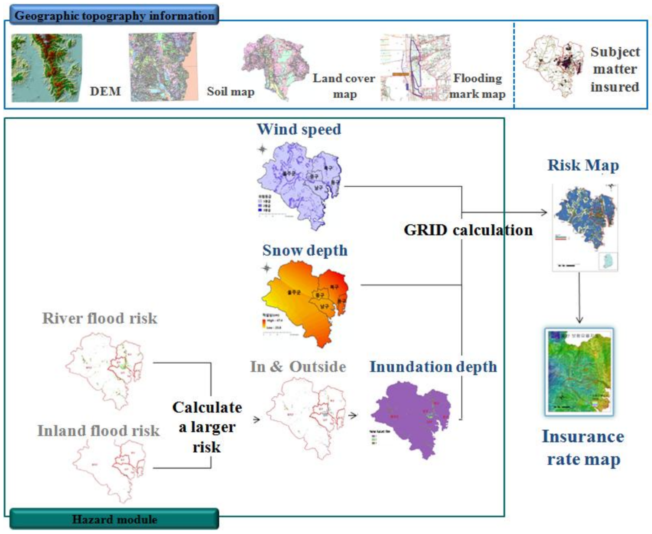
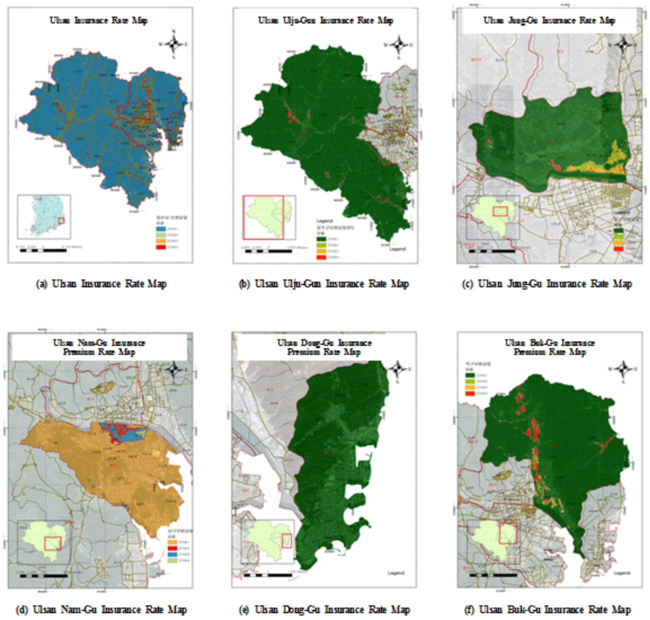


Figure 3. Procedure of producing the storm and flood insurance map.


Ulsan Insurance Rate Map
Ulsan Ulju-Gun Insurance Rate Map
Ulsan Jung-Gu Insurance Rate Map
Ulsan Nam-Gu Insurance Premium Rate Map
Ulsan Dong-Gu Insurance Premium Rate Map
Ulsan Buk-Gu Insurance Premium Rate Map
Legend

(a) Ulsan Insurance Rate Map (b) Ulsan Ulju-Gun Insurance Rate Map (c) Ulsan Jung-Gu Insurance Rate Map

(d) Ulsan Nam-Gu Insurance Rate Map (e) Ulsan Dong-Gu Insurance Rate Map (f) Ulsan Buk-Gu Insurance Rate Map

Figure 4. Storm and flood insurance premium map of Ulsan: (a) whole map, (b) Ulju Gun, (c) Jung Gu, (d) Nam Gu, (e) Dong Gu, (f) Buk Gu

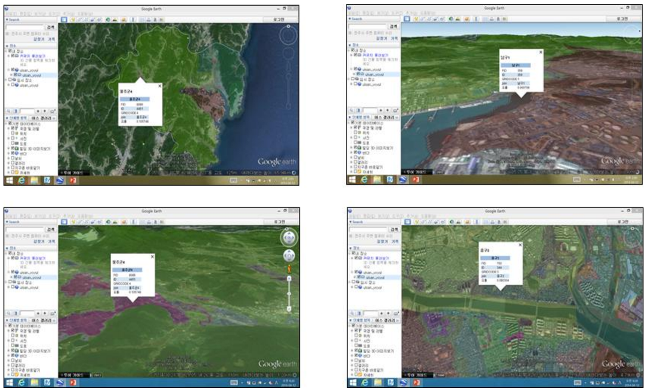


Figure 5. The insurance map in Google earth.

Using GIS for Searching Medieval Texts – Beginnings of a Project

Correlating Digitized Manuscripts and Historical Maps on the Example of Manuscriptorium Digital Library

Tomáš Klimek
National Library of the Czech Republic
Manuscriptorium Digital Library Department
Prague, Czech Republic
Tomas.klimek@nkp.cz

Abstract— **This paper provides a theoretical introduction presenting possibilities and challenges of correlation between medieval texts and historical maps using GIS on the example of Manuscriptorium Digital Library and historical mappings from the area of Central Europe. It discusses methodological questions connected to this task, as well as organizational aspects of concrete suggested solution.**

Keywords- GIS; historical maps; medieval manuscripts; digital libraries of historical documents; information science theory.

I. Introduction

Setting up connections between digitized manuscripts and historical maps can become an important functionality of digital libraries providing historical written documents. It can help users not only to study spatial connections and consequences of historical documents and make easier their representation but it can also provide totally new possibilities for searching and sorting information included in historical texts. The most natural way how to build connections between texts and maps in digital environment is to use GIS.

The aim of this paper is not to describe or project a complete technical solution for connecting digitized manuscripts and historical maps using GIS. It should be rather seen as an introductory theoretical input discussing some methodological and organizational aspects of this task.

The paper will discuss the topic on the concrete example of Manuscriptorium Digital Library and suggest theoretical but realizable solutions. The rest of this paper is organized as follows: Section II introduces with documents we can correlate, Section III introduces with difficulties and challenges of this correlation, Sections IV and V illustrate next practical proceedings and outline of possible final solution.

II. Building material

If we want to correlate digitized historical documents and historical maps, the first question must be – what are our sources? Do we know what to correlate?

A. Textual Documents – Content of Manuscriptorium Digital Library

This paper will discuss the topic using the example of Manuscriptorium Digital Library [1], an international digital library for historical written documents provided and coordinated by the Czech National Library. This is a digital library based upon TEI P5 metadata standards fulfilling the role of an integrator on this field. It provides data of more than 120 institutions across Europe, nowadays metadata for more than 330,000 documents, and more than 25,000 fully digitized items.

The core of the library is created by medieval manuscripts (therefore also its name), but there are also many incunabula, early printed books (printed before 1800) and maps.

B. Maps – First Detailed Mapping and Their Digital Representation

For the purposes of this paper we will need rather to think about the type of available material than to provide its complete overview. Therefore, I will briefly sum up the situation for Central Europe area only.

Concerning the oldest maps detailed enough for our purposes, for countries from Habsburg Empire, we have copies of the “Ist Military Survey“ for our disposal covering the whole area of today´s Czech Republic [2], Austria, Hungary, Slovak Republic, Slovenia and some parts of Rumania, Croatia, Ukraine and Poland [3] in year 1763 and 1787, then we can use Dritte Kursächsische Landesaufnahme for Saxony (1780–1806) [4], and e.g., Dufour Map (1809) for Switzerland [5]. All these mappings have been digitized (some of them several times) and georeferenced, therefore it is a task of cultural policy to set up a cooperation with institutions providing or possessing this data.

III. Difficulties and Challenges

A. Metodological Problems – Different Times of Origin

The first problem of correlation medieval manuscripts and historical maps is obvious. There are no maps in the modern sense of this term coming from medieval period. Although maps as such were known, we cannot speak about

ISBN: 978-1-61208-383-4

modern type of maps – so called “world maps” were being used as a specific representation of theological and cosmological concepts, therefore they were not used for orientation in the landscape and their aim was not to depict and represent the terrain. We know also “itinerary maps” and later also “portulan maps” serving for special purposes only (planning timeline of the journey, navigating along the seashore), and not usable as a modern map at least in inland. It is clear that it would make no sense to use these maps for our correlation – still they do not reflect shapes of terrain in a way which would be sufficient to enable their georeferencing.

The oldest modern maps corresponding enough to the real terrain are not earlier than from the 17th century, and the oldest mappings detailed enough to enable analyses of depicted landscape not earlier than from the second half of the 18th century. This is the time much closer to ours than to the period we are used to call *the Middle Ages* (500-1500). Why to use these documents rather than today’s much more elaborated and detailed maps and map servers?

If you open any book dealing with the history of landscape, settlement, mining, warfare or any other topics demanding the analyses of the form of historical landscape, you can almost always meet with reproductions of mappings from the 18th century although the book deals with much more older history. The reason lies in the fact that these historical mappings depict the landscape before huge changes during the industrial revolution in the 19th century and even bigger changes due to large urbanization and growth of settlements as well as changing agricultural management we experienced during the 20th century.

Nobody would like to claim that there were no big changes in the landscape before the 18th century. We know several processes affecting the form of landscape not less than the processes which followed. For the area of Central Europe I could mention here at least the big colonization in the 13th century, period of destabilization and colder climate leading to abandoning one third of settlements in the first half of the 15th century, and the period of huge depopulation during the Thirty Years War in the first half of the 17th century. However mappings from the 18th century show many aspects, features and facts which have been destroyed and are invisible or hidden in our today’s landscape. Therefore, supposing the knowledge of basic changes, they can be efficiently used by scholars to study older situations.

B. Practical Problems

Dealing with medieval texts and geographical locations we have to challenge several specific difficulties. Apart from modern texts we do not have complete indexes for localities due to two facts. Authors and scribes of medieval texts used many various forms of local names for concrete settlements. Therefore we need special dictionaries providing synonyms and antonyms for older local names.[3] For many cases we will even need to look into specialized monographs dealing with some concrete situations.

At the same time we have here many locations which do not already exist. Sometimes we even do not know its original position in the landscape. Sometimes we will be able to use localizations of such places from historiographical or archeological literature, sometimes we will be forced to exclude some localities from the correlation.

Another difficulty will be caused by variability of information types in Manuscriptorium Digital Library. We must imagine that only a small part of included information is in a text form. Images of digitized documents are of course not searchable. Then we have here so called metadata including technical metadata and descriptions, and then other types of representation - full text editions correlated to images of original documents. We cannot distinguish strictly between descriptions and full text editions parts of descriptions used for identifying the texts are taken directly from texts from original documents (so called incipits, explicits, etc.). These parts are rather data than metadata. On the other hand, other parts of descriptions refer to holders of physical documents, history of concrete items, and their physical features like origin of the writing material, type of bindings, etc. These both totally different types of information have their own value and purpose speaking about their correlation with maps, and must be sorted separately.

Another issue is caused by various coordinate systems of historical maps. However for the case of Manuscriptorium’s texts we can use maps digitized by other institutions (e.g., The University of Jan Evangelista Purkyně in Ústí nad Labem – Faculty of Environment; Research Institute of Geodesy, Topography and Cartography in Zdiby) which have already converted the old coordinate systems of some maps and georeferenced them to modern maps.

Concerning belongings of concrete places to various countries, considering the changing borders throughout the time, to avoid confusions, we should come out from the current situation.

IV. PRACTICAL STEPS DO BE DONE

Apart from starting negotiations with institutions holding data concerning historical mappings, we would have to enrich existing data of the digital library of historical written documents by XML tagging providing geographical coordinates. The reason why to use XML and e.g,. not to start with a special ontology is following – surely we can use existing thesauri or dictionaries like CERL (gathering local and personal names from early printed books) for this work and try to semi automatize the process but medieval texts are unique concerning the issue of local names. [6] Many of them contain many mentions about abandoned and forgotten localities, we can meet with many variants of names and many names used for more localities, etc. therefore each text demands special care and linking to universal dictionary becomes extremely problematic. Of course separate internal ontology using RDF can be created as a usable output – maybe it can appear as necessary because of the need of representing relations between localities and their hierarchical levels; nevertheless also this internal ontology would have to be linked to concrete texts by all means and probably won´t be transferable.

For this activity there is a good chance to use crowdsourcing. As other project show, work with maps is attractive enough to arouse interest. In this case it would be essential to develop a user-friendly editorial tool for XML tagging providing a workspace with a modern map open able with tagged text. Users would then just click on a word from the text and to the map to add necessary information; the tool would prepare all the technical work automatically.

V. Outline of possible solution

The final result of our effort would enhance facilities of digital research environment of Manuscriptorium. Users could easily work with geographical data included in provided documents; they would be able to study or represent this data on various time layers maps which would enable them to study some mass phenomena like spreading information across the space, setting up events to the landscape in the literature (representation could reflect the sequence of localities mentioned in the text), etc., etc. At the same time – a map could be incorporated into searching of the digital library – the user could choose one or more localities or determine an area and the searching tool linked to the map environment would narrow the next user query (i.e. the system would look for documents with metadata containing at least one of localities included in the query and possible in defined fields – tags only). Processing texts in this way would also increase the searchability of maps at the same time because maps would be connected to metadata of digitized written documents. All this would again substantially increase the usability of the digital library for research.

References

[1] Manuscriptorium. [Online]. Available from: http://www.manuscriptorium.com/ 2014.10.09

[2] Presentation of old maps covering the area of Czechia, Moravia and Silesia. [Online]. Available from: http://oldmaps.geolab.cz/ 2014.10.09.

[3] Historical Maps of the Habsburg Empire. [Online]. Available from: http://mapire.eu/en/ 2014.10.09.

[4] Geoportal of Hochschule für Technik und Wirtschaft Dresden - Geoportal of Faculty of Spatial Information. [Online]. Available from: http://geoinformatik.htw-dresden.de/ 2014.10.09.

[5] Dufour Map of Switzerland. [Online]. Available from: http://map.geo.admin.ch/ 2014.10.09.

[6] CERL Thesaurus. [Online]. Available from: http://www.cerl.org/ 2014.10.09.

ISBN: 978-1-61208-383-4

Creating Knowledge-based Dynamical Visualisation and Computation

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU),
Leibniz Universität Hannover,
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

***Abstract*—The research conducted in this paper presents an implementation for creating knowledge-based dynamical visualisation and computation. The core knowledge is based on long-term knowledge resources further developed for several decades and used with many applications scenarios utilising multi-disciplinary and multi-lingual content and context like references, associations, and knowledge container collections. A major goal of the application case studies shown here is creating context-sensitive dynamical program components and algorithms from selected knowledge content. The selections are results of dynamical workflows, which are part of component implementations, e.g., including search processes and result matrix generators. Previous research has shown that long-term knowledge resources are the most important and most valuable component of long-term approaches and solutions. Here, the structures and classifications are used with independent database frameworks and programming interfaces. The results show that the methodological foundations and knowledge resources are very well suitable as long-term core base, as well as for creating dynamical application components, e.g., for visualisation and computation in multi-disciplinary, geoscientific, and spatial context. The knowledge resources can refer to any kind of resources. The overall environment allows to develop and govern extensive content structures and promote their long-term vitality.**

Keywords*–Advanced Knowledge Discovery; Universal Decimal Classification; Conceptual Knowledge; Dynamical Visualisation and Computation; Geoscientific Applications.

I. INTRODUCTION

Long-term Knowledge resources can be created and used for universal documentation and re-use of content. The re-use includes discovery, as well as gathering new results and creating new applications. The knowledge resources [1] can refer to any kind of resources, e.g., to natural sciences resources or historical geographic resources [2]. Basics of knowledge organisation [3] and multi-lingual lexical linked data [4] have been discussed for various disciplines and shown the huge potential and value of the knowledge. This also shows the benefits of linking with universal classifications, especially with consequently numerical notations, which can be easily and most flexibly and efficiently used with modern applications components. Further, on the one hand, information services benefit from a comprehensive and holistic model for evaluation [5] and on the other hand, they align with the benefits for a quality management of information services [6].

The paper presents a new implementation for creating knowledge-based dynamical visualisation and computation, which have not been integrated before for that purpose. Therefore, a major goal of the application case studies is creating dynamical program components and algorithms based on knowledge resources. The different previous projects and case studies have already shown that the combination of knowledge resources with integrated conceptual knowledge references can be used for the creation of dynamical applications.

The dynamical visualisation and computation based on knowledge resources does have numberless applications. Some prominent examples with the research presented here are knowledge discovery, visualising result matrices from workflows or search processes, and creating objects and extending knowledge resources. The framework presented here is a high level framework interconnecting several frameworks for complex system architecture, multi-column operation, and long-term creation for main resources. Therefore, the required approach is considered to be necessarily most complex from knowledge and implementation point of view.

Following the Geo Exploration and Information case studies [7] based on the actmap framework [8] a number of developments for the deployment of High End Computing resources and technologies with integrated systems are still state of the art. In addition, including the structural and conceptual knowledge based on the knowledge resources, research has been done for a different special database framework, which is as well autonomous and can be used for the creation of standalone dynamical and portable application components. The components can be integrated with the existing frameworks, as well as they can be used as standalone interactive applications.

This research shows details of the latest case studies and discusses the up-to-date experiences from the implementation of the dynamical components and their integration with knowledge resources' structures and workflows.

This paper is organised as follows. Section II introduces the methodological bases, Section III presents the fundamental implementations based on knowledge resources and computation, discussing the foundations, architecture, framework, integration, and dynamical visualisation and computation. Section IV shows the details of the implementation and the resulting components, from geo sets, computation to index selection and some views from the resulting visualisations. Sections V and VI evaluate the main results and summarise the lessons learned, conclusions and future work.

II. STATE-OF-THE-ART AND MOTIVATION

The creation of long-term knowledge resources and utilisation methods is one of the most pressing goals in information science as the masses of data and the loss of knowledge in society are steadily increasing in all areas. Existing projects employ segment-like spectra of disciplines in their focus. Examples are large digital libraries and projects like the Europeana [9] and the World Digital Library (WDL) [10].

As existing projects, e.g., which are only concentrating on bibliographic means, do not focus on such integration, use different and mostly isolated classifications and schemes for different areas and specific purposes. For example, there is a small number of general classifications, which are mostly used in library context. Although such classifications are used in many thousands of institutions worldwide this is neither a general use case or application scenario nor a significant share of the overall knowledge. They are missing to provide facilities for arbitrary kinds of objects, e.g., factual data and trans-disciplinary context in information science and natural sciences.

In contrast to that, the state of the art for documentation of universal, conceptual knowledge is the Universal Decimal Classification (UDC) [11], which is one of the very few classifications providing a universal classification. Besides public interfaces, the implementations of the known application scenarios are not publicly available in common. Anyhow, all known scenarios have in common that they deploy only a small subset of available classifications and in the vast majority the classification process is not automated. It is necessary to develop logical structures in order to govern the existing big data today and in future, especially in volume, variability, and velocity. This is necessary in order to keep the information addressable and maintain the quality of data on long-term.

Beyond the focus of the mentioned segment-like projects the knowledge resources and concepts discussed and implemented in this research focus on the trans-disciplinary integration of arbitrary different segments and disciplines and a universal usability based on factual data and criteria. The documentation and context also integrates and refers to content and context, e.g., conceptual, procedural, and metacognitive knowledge and allows for a huge range of possible scenarios. This is a driving force to extend the use of classification in trans- and multi-disciplinary context and transfer the experiences from deploying a classification for long-term documentation and application.

III. Methodological bases: Logical structures

To work on that goal requires to define information units and to care for depositing an appropriate segmentation in sub-units. The information units require links to the related units, e.g., superunits. The challenges are to define these structures and units for data used in different disciplinary context, in one discipline, as well as in multi-disciplinary context. These logical structures are the basic precondition for the development of functioning algorithms, which can access the units and whose application can be perfected in a next self-learning step. The tries of using unstructured data result in the fact that data volumes, variabilities, and volumes devaluate the resulting values of requests. Any isolated technological approaches to the big data challenge have shown not to be constructive. A sustainable approach has to consider the data and structure itself.

- The first step is the preconditional definition of a logical, commonly valid structure for the data.
- The second step is the planning for the applications based on the logical structures in step one.
- The third step is the creation of algorithms regarding the data and data retrieval, interfaces, and workflows based on steps one and two. The fourth step is the planning of the implementation. This includes data format, platforms, and applications.

Further, the creation, development, and operation of the content and components require to consider and define the essential plans, especially:

- Plans for extending structures.
- Preparations for all required interfaces for the newly extended structures should be done.
- Plans for self-learning components.
- Plans for container formats and utilisation.
- Plans for sustainable handling of data lifecycles, data formats, and system resources,

The early stage of planning requires a concept catalogue. So far, the activities are commonly in a pre-planning stage. The next step includes case studies on structures, algorithms, and collaborational issues (efficiency, economical cycle).

IV. Implementation: Knowledge and computation

The implementation shows the context-sensitive dynamical components based on the knowledge resources. The knowledge resources provide the structure and integrate the factual knowledge, the references, including the references for the object classification views required for the dynamic utilisation, ensuring integration [12] and sustainability [13].

Previous case studies have shown that any suitable cartographic visualisation can be used for the presentation of the results, for example, with the Generic Mapping Tools [14] (e.g., filtering, trend fitting, gridding, views, and projections) or creating exports and imports with various products. Most available cartographic visualisation products are too specialised in order to handle advanced knowledge workflows one the one hand and dynamical results on cognitive context on the other hand. In the presented case where the application should concentrate on the intention of presenting a special result in an abstract way we require special and flexible facilities for dynamical sketch drawings. The more, in the special case the cognitive background forbids to concentrate on detailed cartographic visualisation or mixing with modern ways of geographic conventions. Historical names, locations, and context are not adequately represented by existing modern frameworks.

Regarding both requirements for this study are fully complied by the flexibility of the implementation. The knowledge resources themselves are not restrictive regarding the use of other components for other purposes.

A. Implementation foundations

The implementation for dynamical visualisation and computation is based on the framework for the architecture for documentation and development of advanced scientific computing and multi-disciplinary knowledge [15]. The architecture implemented for an economical long-term strategy is based on different development blocks. Figure 1 shows the three main columns: Application resources, knowledge resources, and originary resources. The central block in the "Collaboration house" framework architecture [16], are the knowledge resources, scientific resources, databases, containers, and documentation (e.g., LX [1], databases, containers, list resources). These can be based on and refer to the originary resources and sources (photos, scientific data, literature). The knowledge resources are used as a universal component for compute and storage workflows. Application resources and components (Active Source, Active Map, local applications) are implementations for analysing, utilising, and processing data and making the information and knowledge accessible. The related information, all data, and algorithm objects presented are copyright the author of this paper, LX Foundation Scientific Resources [1], all rights reserved.

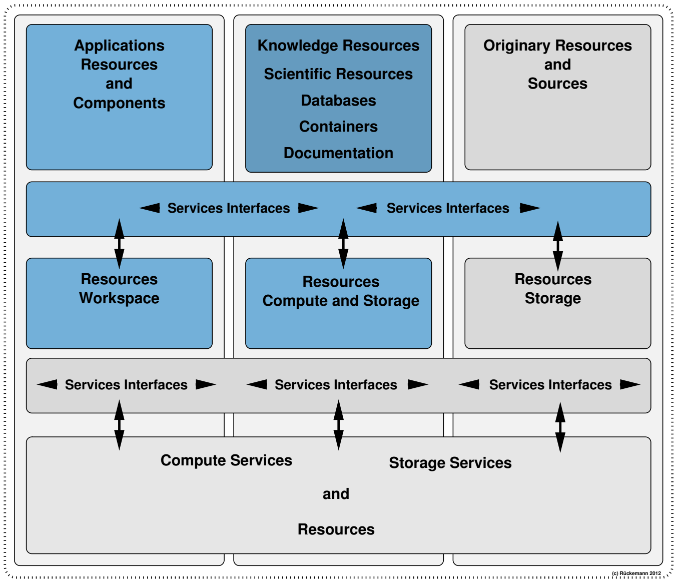


Figure 1. Architecture: Columns of practical dimensions. The knowledge resources are the central component within the long-term architecture.

The LX structure and the classification references based on UDC [11], especially mentioning the well structured editions [17] and the multi-lingual features [18], are essential means for the processing workflows and evaluation of the knowledge objects and containers. Both provide strong multi-disciplinary and multi-lingual support. The three blocks are supported by services' interfaces. The interfaces interact with the physical resources, in the local workspace, in the compute and storage resources the knowledge resources are situated, and in the storage resources for the originary resources. All of these do allow for advanced scientific computing and data processing, as well as the access of compute and storage resources via services interfaces. The resources' needs depend on the application scenarios to be implemented for user groups.

B. Implementation and integration

The context of the application components is fully integrated with the knowledge resources and dynamical components [19]. The screenshot (Figure 2) illustrates some features.

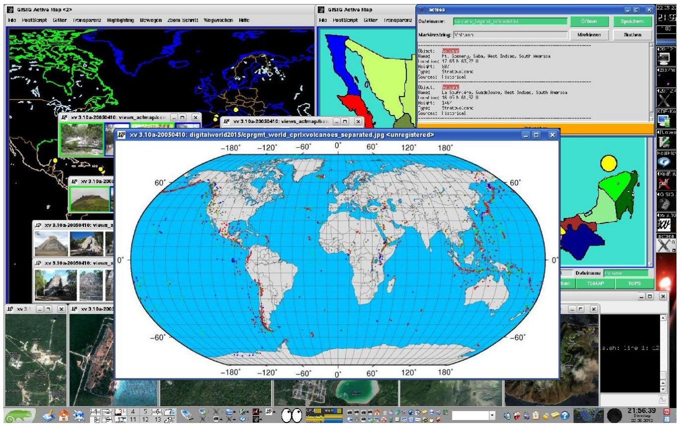


Figure 2. Dynamical use of information systems and scientific computing with multi-disciplinary and universal knowledge resources [16].

Shown examples illustrate features of Active Source, computed and filtered views, LX information, and aerial site photographs, e.g., from Google Maps. Many general aspects of dynamical use of information systems and scientific computing have been analysed with the collaboration house case studies.

C. Implemented content and application dynamics

The main groups of challenges are resulting from the content and from the applications. From content side: The knowledge resources provide the central repository and infrastructure ('Knowledge as a Service') for discovery and component creation. From application side: The dynamical components can deploy the resources to any extent and in any step of workflows.

Any part of the components and features can be assembled from the knowledge resources' workflows. As an example, the site context and factual data components, the database and the graphical user interface components including event definition and management can be dynamically created via transform routines and `concatenate` operations.

D. Dynamical visualisation and computation implementation

A number of different visualisation tools and frameworks have been analysed in the latest case studies. The results presented here were mostly realised with Tcl/Tk [20] for the dynamical visualisations, Fortran and C based programs for required algorithms, Message Passing Interface (MPI) [21], and Perl for dynamical scripting.

Many components have been developed for the actmap framework [8] and successfully used and verified in context with existing scenarios. Besides the actmap framework additional possibilities of creating application programming interfaces and graphical user interfaces for dynamical visualisation of knowledge matrices have to be analysed.

As a simple example for a dynamical, portable, and standalone component an application like Tclworld has been considered [22]. The application is built on a very portable Tool Command Language (TCL) base integrating programming, database, and user interfaces.

The database application programming interface [23] is very simple, portable, and extendable. The database graphical user interface [24] can be used within the same application and is based on a rapid prototyping concept. Both interface models allow dynamical control and extension of any features regarding the application, as well as for the content.

V. Implementation and resulting components

Creating knowledge-based dynamical components starts at the application to knowledge resources' interface level. The conventional components and processes have been described and discussed in practice in detail in previous research, regarding Active Map Software [8], case studies [7], and knowledge integration for classification and computation [12].

For the visualisation and computation using the matrix objects with spatial and georeferenced context, a new application instance "lxworlddynamic" has been created based on the knowledge resources and interfaces. This component re-using the Tclworld interfaces is the required extension complementary to the actmap framework components. All parts of the component shall

- integrate with the knowledge resources and existing components, e.g., georeferenced objects,
- be usable interactively,
- have access to the content, e.g., index selection,
- facilitate a standalone application assembly, and
- allow a flexible configuration of all aspects of the applications' visualisation and computation.

Especially, the component requires a site handler database, has to refer to sets of georeferenced objects, a level handler for selecting levels of detail, to create groups of matrix objects, to support for an individualised configuration, to facilitate index selections on the generated matrix elements, to visualise the matrix elements and context graphically, and has to provide associated data textually and numerically.

A. Knowledge resources and geo sets

The integrated information systems can generate result matrices based on the available components and workflows. The result matrix generators can be configured to deliver any kind of result matrix. One base for the implementation is the generation of georeferences data resulting from requests. The listing (Figure 3) shows an example of a result matrix, an excerpt of the generated site handler database.

```
{LX Site 20.687652 -88.567674} site 20.687652 88.567674
{LX Site 20.682658 -88.570147} site 20.682658 88.570147
{LX Site 20.682859 -88.568548} site 20.682859 88.568548
{LX Site 21.210859 -86.80352} site 21.210859 86.80352
{LX Site 21.097633 -86.796799} site 21.097633 86.796799
{LX Site 21.157199 -86.834736} site 21.157199 86.834736
{LX Site 21.157199 -86.834736} site 21.157199 86.834736
{LX Site 21.094751 -86.812248} site 21.094751 86.812248
{LX Site 41.377968 2.17804} site 41.377968 -2.17804
{LX Site 41.375842 2.177696} site 41.375842 -2.177696
{LX Site 38.676439 -0.198618} site 38.676439 0.198618
{LX Site 38.677683 -0.198103} site 38.677683 0.198103
{LX Site 21.234502 -86.740494} site 21.234502 86.740494
{LX Site 21.184412 -86.807528} site 21.184412 86.807528
{LX Site 16.043421 -61.663857} site 16.043421 61.663857
{LX Site 16.043153 -61.663374} site 16.043153 61.663374
{LX Site 17.633225 -63.236961} site 17.633225 63.236961
{LX Site 17.633225 -63.236961} site 17.633225 63.236961
{LX Site 51.151786 10.415039} site 51.151786 -10.415039
{LX Site 20.214301 -87.429103} site 20.214301 87.429103
{LX Site 20.493276 -87.735701} site 20.493276 87.735701
{LX Site 20.494663 -87.720294} site 20.494663 87.720294
{LX Site 20.494761 -87.720138} site 20.494761 87.720138
{LX Site 40.821961 14.428868} site 40.821961 -14.428868
{LX Site 20.365228 -87.452545} site 20.365228 87.452545
{LX Site 20.365228 -87.452545} site 20.365228 87.452545
```

Figure 3. Excerpt of generated site handler database (lxworlddynamic).

The database is the result of a request summarising results on objects referring to a defined context, in this case references between archaeological and geological objects.

The framework provides a number of features like level handlers and sets of object georeferences. The listing (Figure 4) shows an excerpt of the generated site level handler.

```
foreach i {
  { ... }
  { ... }
  ...
} {+ $i level 2}
```

Figure 4. Excerpt of generated site level handler (lxworlddynamic).

The level handler manages the site handler database, which can also be generated and updated dynamically. Appropriate entries are managed by the geo::set. The listing (Figure 5) shows an excerpt of the generated geo set.

```
geo::Set {
  { ... } site ... ...
  { ... } site ... ...
  ...
}
```

Figure 5. Excerpt of generated geo::set (lxworlddynamic).

Groups of objects, e.g., associated archaeological, geological, meteorite, and volcanological sites, as well as subgroups like pottery and stones, can be dynamically associated and handled in the generated component. The listing (Figure 6) shows an excerpt of the generated database matrix.

```
 + {Archaeological site} : {A selected site with findings
    of human activity, complementary to a
    {Geological site} .
    These sites have been dynamically created
    from a request to the LX knowledge resources ... .
    These sites have been ... .
    }

 + {Geological site} : {A selected site with geological
    findings, e.g., a {Volcanological site} or a
```

```
   {Meteorite site} , complementary to an
   {Archaeological site} .
   This site has been ...}

+ {Metorite site} : {A selected site with meteorite
   findings, e.g., meteorite crater, a special
   {Geological site} .
   This site has been ...}

+ {Volcanological site} : {A selected site with
   volcanological findings, e.g., volcanological
   findings like a volcano or fumarole, a special
   {Geological site} .
   This site has been ...}

+ Pottery    : {Archaeological site} major.countries {
 Italy France Spain Greece}

+ Stone      : {Geological site} major.countries {Italy
 France Spain Greece}
```

Figure 6. Excerpt of generated database matrix (lxworlddynamic).

Here, the matrix includes site and object types for the respective matrix with excerpts of descriptions and linked references.

B. Index selection and configuration

When a representation of matrix objects in dynamical spatial cartographic context is possible then selected objects can be integrated either from the matrix elements (e.g., sites) or from the context elements and references (e.g., cities, mounts, lakes, roads, rails, rivers, and grids). The listing (Figure 7) shows an excerpt of the generated index selection.

```
foreach i [geo::Names] {
   if {[lindex $geo::db($i) 0]=="city"} {
       "LX_World_database" $i : {city} loc [lrange $
        geo::db($i) 1 end]
   }
   if {[lindex $geo::db($i) 0]=="mount"} {
       "LX_World_database" $i : {mount} loc [lrange $
        geo::db($i) 1 end]
   }
   if {[lindex $geo::db($i) 0]=="site"} {
       "LX_World_database" $i : {site} loc [lrange $
        geo::db($i) 1 end]
   }
   if {[lindex $geo::db($i) 0]=="lake"} {
       "LX_World_database" $i : {lake} loc [lrange $
        geo::db($i) 1 end]
   }
   if {[lindex $geo::db($i) 0]=="road"} {
       "LX_World_database" $i : {road} loc [lrange $
        geo::db($i) 1 end]
   }
   if {[lindex $geo::db($i) 0]=="rail"} {
       "LX_World_database" $i : {rail} loc [lrange $
        geo::db($i) 1 end]
   }
   if {[lindex $geo::db($i) 0]=="river"} {
       "LX_World_database" $i : {river} loc [lrange $
        geo::db($i) 1 end]
   }
   if {[lindex $geo::db($i) 0]=="grid"} {
       "LX_World_database" $i : {grid} loc [lrange $
        geo::db($i) 1 end]
   }
}
```

Figure 7. Excerpt of generated index selection (lxworlddynamic).

Any part of the dynamically generated components can be individualised depending on context-sensitive attributes and workflow configuration. The listing (Figure 8) shows an example for the generated on-the-fly-symbol used for “sites”.

```
set bitmaps(site) [image create bitmap -data [strimj::xbm
  "
....#....
...###...
..#...#..
.#.###.#.
##.###.##
.#.###.#.
..#...#..
...###...
....#...."] -foreground darkviolet]
```

Figure 8. Excerpt of generated on-the-fly symbol for sites (lxworlddynamic).

Different symbols can be integrated for different sites or for different groups. The objects with their symbols are only visible if the defined level, which is handled by the level handler, is active in the interactive view.

C. Dynamical visualisation and computation

The following image (Figure 9) shows a screenshot of a resulting dynamical visualisation of items in the result matrix, in this case the resulting archaeological context sites. The generated application utilises all the features so far described with the implementation.

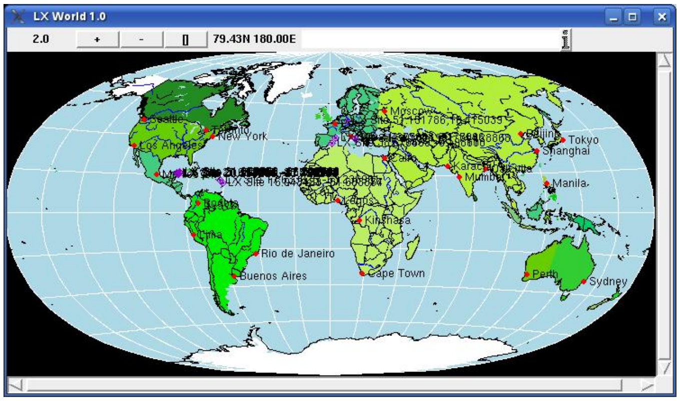


Figure 9. Archaeological context sites in interactive, dynamically generated spatial application (lxworlddynamic).

The screenshot illustrates the dynamical visualisation of the matrix elements for the context of the respective results. With the workflow a spatial context has been chosen for the matrix, creating the components. The spatial application component has been assembled by the workflow, integrating the object item references from the result matrix and secondary information from the referring knowledge resources objects with a dynamical and interactive view of the matrix.

Figure 10 shows a screenshot of a resulting dynamical zoom visualisation of matrix results and secondary information on geological and archaeological context sites. The partially shown superpositioning effect of the respective zoom is still visible in order to show the results, which can be separated in different cognitive views, different zooms, and event sensitive actions. The implementor can do anything with this feature he is interested in, e.g., use the interactive features for label stacks and level effects, being sensitive for single results or result group. Workflow sensitive cartographic material objects (e.g., cities, land, sea, countries, border lines, grid lines) are shown for orientation and context and support cognitive feature display. The shown zoom value and the scroll bars indicate the level of detail.

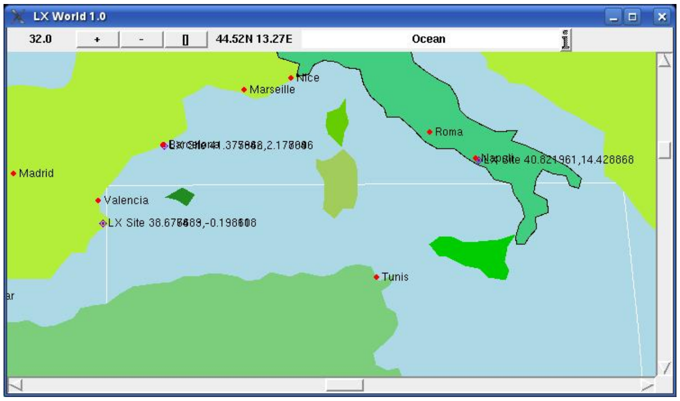


Figure 10. Zoom archaeological and geological sites and context, integrated in interactive, dynamically generated spatial application (lxworlddynamic).

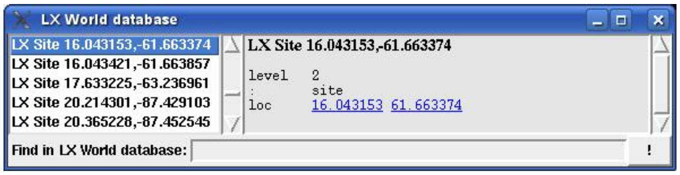
The screenshot illustrates the matrix elements and their references within the knowledge resources. In this example, an active context-sensitive window component is delivering the secondary information. This component is actively communicating with the other components. The site entity data regarding levels and objects referred from the matrix elements is dynamically available (Figure 11).

```
LX Site 16.043153,-61.663374

level   2
:       site
loc     16.043153 61.663374
```

Figure 11. Example of a single site entity data extract (lxworlddynamic).

Here the matrix elements are referring to the attributes of the knowledge resources' objects, e.g., sites and cities. In this example, the displayed data excerpt includes the level, the type, and the location of the respective site. Figure 12 shows a screenshot of the corresponding dynamical site database.



```
LX Site 16.043153,-61.663374
LX Site 16.043421,-61.663857
LX Site 17.633225,-63.236961
LX Site 20.214301,-87.429103
LX Site 20.365228,-87.452545
```

```
LX Site 16.043153,-61.663374

level   2
:       site
loc     16.043153 61.663374
```

Find in LX World database:

Figure 12. Site database, secondary information (lxworlddynamic).

The site database can be accessed by integrated or external applications' components, e.g., for searches, references or for generating further result matrices. The screenshot illustrates the matrix elements and their references within the knowledge resources. In this example, an active context-sensitive window is delivering the secondary information.

VI. Evaluation

Application components can be created and assembled dynamically from any workflow. Knowledge objects can be used efficiently with any dynamical components. Database interfaces can be used dynamically and efficiently with the components. Graphical User Interfaces (GUI) and Application Programming Interfaces (API) can be used most flexible on that dynamical base.

The information created from arbitrary numbers of resources' objects in this excerpt includes site labels, level information, category information, as well as the georeferences. The components can trigger any instances and events dynamically and interactively. This allows any kind of processing, computation, and visualisation, from sketch like visualisation to special cartographical mapping. All the more, the dynamical components based on the knowledge resources and the lxworlddynamic frameworks allow for the interactive, autonomous components' generation. The components have been successfully implemented on a number of operating systems (e.g., SuSE Linux, Debian, Red Hat, and Scientific Linux, as well with older and up-to-date distributions).

The knowledge and system architecture allows to seamlessly integrate with all the steps required for a sustainable implementation of the methodological bases. The case studies done over several decades of knowledge resources' creation and development and two decades of application component developments have shown that plans for extending structures, creating interfaces, self-learning components, container formats, and integration with life cycles and systems resources' operation can be assured even for long-term application.

VII. Conclusions

It has been demonstrated that with the proposed framework and concept context-sensitive dynamical components can be successfully created on base of universal knowledge resources. It has been shown that even standalone dynamical components can be created based on the implementation of the foundations, supporting arbitrary dynamical, modular, portable, and extendable database application programming and database graphical user interfaces. The implementation can utilise workflows and algorithms for knowledge discovery and selection up to intelligent application component creation. It integrates very efficiently in workflow chains, e.g., for computation and visualisation, and is very well usable even for rapid prototyping environments.

The integration is non-invasive regarding the knowledge resources for uni-directional visualisation and computation. If the intention with an application scenario is to update information consistently then multi-directional workflows can also update objects in side knowledge resources or containers from the created components in arbitrary ways, ensuring consistency and plausibility, as well as following management and security policies. Future work will be focussed on issues and usability beyond the plain Big Data approaches, resulting from data values, creating implementations supporting data vitality.

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. Special thanks go to the scientific colleagues at the Gottfried Wilhelm Leibniz Bibliothek (GWLB) Hannover, especially Dr. Friedrich Hülsmann, for prolific discussion within the long-term "Knowledge in Motion" (KiM) project, for inspiration, and practical case studies. Many thanks go to Mrs. Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, to the Institute for Legal Informatics (IRI), Leibniz Universität Hannover, and to the scientific colleagues at the Westfälische Wilhelms-Universität (WWU), for discussion, support, and sharing experiences on collaborative computing and knowledge resources and for participating in fruitful case studies, as well as to the participants of the postgraduate European Legal Informatics Study Programme (EULISP) for prolific discussion of scientific, legal, and technical aspects over the last years. Last but not least we thank Richard Suchenwirth for his initial implementation of Tclworld, which could be used as a tool for handling result matrices.

REFERENCES

[1] "LX-Project," 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/#LX (Information) [accessed: 2014-10-05].

[2] E. Dodsworth and L. W. Laliberte, Eds., Discovering and using historical geographic resources on the Web: A practical guide for Librarians. Lanham: Rowman and Littlefield, 2014, ISBN: 0-8108-914-1.

[3] S. A. Keller and R. Schneider, Eds., Wissensorganisation und -repräsentation mit digitalen Technologien. Walter de Gruyter GmbH, 2014, Bibliotheks- und Informationspraxis, ISSN: 0179-0986, Band 55, ISBN: 3-11-031270-0.

[4] E. W. De Luca and I. Dahlberg, "Die Multilingual Lexical Linked Data Cloud: Eine mögliche Zugangsoptimierung?" Information Wissenschaft & Praxis, vol. 65, no. 4–5, 2014, pp. 279–287, Deutsche Gesellschaft für Information und Wissen e.V. (DGI), Ed., De Gruyter Saur, ISSN: 1434-4653, e-ISSN: 1619-4292, DOI: 10.1515/iwp-2014-0040, (title in English: The Multilingual Lexical Linked Data Cloud: A possible semantic-based access to the Web?).

[5] L. Schumann and W. G. Stock, "Ein umfassendes ganzheitliches Modell für Evaluation und Akzeptanzanalysen von Informationsdiensten: Das Information Service Evaluation (ISE) Modell," Information Wissenschaft & Praxis, vol. 65, no. 4–5, 2014, pp. 239–246, Deutsche Gesellschaft für Information und Wissen e.V. (DGI), Ed., De Gruyter Saur, ISSN: 1434-4653, e-ISSN: 1619-4292, DOI: 10.1515/iwp-2014-0043, (title in English: A comprehensive holistic model for evaluation and acceptance analyses of information services: The Information Service Evaluation (ISE) model).

[6] G. Isaew and A. Roganow, "Qualitätssteuerung von Informationssystemen: Theoretisch-methodologische Grundlagen," Information Wissenschaft & Praxis, vol. 65, no. 4–5, 2014, pp. 271–278, Deutsche Gesellschaft für Information und Wissen e.V. (DGI), Ed., De Gruyter Saur, ISSN: 1434-4653, e-ISSN: 1619-4292 DOI: 10.1515/iwp-2014-0044, (title in English: Quality Management of Information Systems: Theoretical and methodological basics).

[7] "Geo Exploration and Information (GEXI)," 1996, 1999, 2010, 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/index.html#GEXI [accessed: 2014-10-05].

[8] C.-P. Rückemann, "Active Map Software," 2001, 2005, 2012, URL: http://wwwmath.uni-muenster.de/cs/u/ruckema (information, data, abstract) [accessed: 2012-01-01].

[9] "Europeana," 2014, URL: http://www.europeana.eu/ [accessed: 2014-11-30].

[10] "WDL, World Digital Library," 2014, URL: http://www.wdl.org [accessed: 2014-11-30].

[11] UDC, Universal Decimal Classification. British Standards Institute (BSI), 2005, complete Edition, ISBN: 0-580-45482-7, Vol. 1 and 2.

[12] C.-P. Rückemann, "Knowledge Integration for Scientific Classification and Computation," in The Fourth Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of The 12th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), September 22–28, 2014, Rhodes, Greece, Proceedings of the American Institute of Physics (AIP), AIP Conference Proceedings. AIP Press, 2014, ISSN: 0094-243X.

[13] C.-P. Rückemann, "Long-term Sustainable Knowledge Classification with Scientific Computing: The Multi-disciplinary View on Natural Sciences and Humanities," International Journal on Advances in Software, vol. 7, no. 1&2, 2014, pp. 302–317, ISSN: 1942-2628.

[14] "GMT - Generic Mapping Tools," 2014, URL: http://imina.soest.hawaii.edu/gmt [accessed: 2014-01-12].

[15] C.-P. Rückemann, "High End Computing Using Advanced Archaeology and Geoscience Objects," International Journal On Advances in Intelligent Systems, vol. 6, no. 3&4, 2013, pp. 235–255, ISSN: 1942-2679, LCCN: 2008212456 (Library of Congress), URL: http://www.iariajournals.org/intelligent_systems/intsys_v6_n34_2013_paged.pdf [accessed: 2014-11-30].

[16] C.-P. Rückemann, "Enabling Dynamical Use of Integrated Systems and Scientific Supercomputing Resources for Archaeological Information Systems," in Proc. INFOCOMP 2012, Oct. 21–26, 2012, Venice, Italy, 2012, pp. 36–41, ISBN: 978-1-61208-226-4.

[17] "Multilingual Universal Decimal Classification Summary," 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php [accessed: 2014-11-30].

[18] "UDC Online," 2014, http://www.udc-hub.com/ [accessed: 2014-10-05].

[19] 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, 2013, pp. 200–213, ISSN: 1942-261x.

[20] "Tcl Developer Site," 2014, URL: http://dev.scriptics.com/ [accessed: 2014-11-30].

[21] "Open-MPI," 2014, URL: http://www.open-mpi.org [acc.: 2014-06-29].

[22] R. Suchenwirth, "Tclworld, Version 0.6, Konstanz, 2002-01-31," 2002, URL: http://wiki.tcl.tk/_repo/wiki_images/tm06.zip [acc.: 2014-11-30].

[23] R. Suchenwirth, "dbapi – A simple database API," 2002, URL: http://wiki.tcl.tk/_repo/wiki_images/tm06.zip [accessed: 2014-11-30].

[24] R. Suchenwirth, "dbgui.tcl – A little database GUI," 2002, URL: http://wiki.tcl.tk/_repo/wiki_images/tm06.zip [accessed: 2014-11-30].

Geographic Data Modeling for NoSQL Document-Oriented Databases

Wagner Boaventura Filho, Harley Vera Olivera, Maristela Holanda, Aleteia A. Favacho

Department of Computer Science
University of Brasília
Brasília, Brazil
{wagnerbf, harleyve}@gmail.com, {mholanda, aleteia}@cic.unb.br

***Abstract*—The integration of Geographical Information Systems (GIS) with relational databases (RDBMS) and the search for interoperable standards among geospatial systems have been featured focuses on the agendas of academia, industry, and the spatial data user community in for some time now. Subsequently, in database technologies, some of the new issues increasingly debated are non-conventional applications, including NoSQL (Not only SQL) databases, which were initially created in response to the needs for better scalability, lower latency and higher flexibility in an era of bigdata and cloud computing. These non-functional aspects, which are very common in the treatment of spatial data, are the main reason for using NoSQL database. However, currently there are no systematic studies on the data modeling for NoSQL databases, especially the document-oriented ones. Therefore, this article proposes a NoSQL data modeling standard, introducing modeling techniques that can be used on document-oriented databases, including geographical features. In addition, to validate the proposed model, a study case was implemented using geographic information on changes in the land use of Brazilian biomes.**

***Keywords*–NoSQL; GIS; Data modeling; Document-oriented database.**

I. Introduction

Huge amounts of spatial data are produced daily. They are generated by satellites, telescopes, sensor networks, and provide information that is growing exponentially. The management of this data is currently performed in most cases by relational databases with spatial extensions that provide centralized control of data, redundancy control and elimination of inconsistencies [1]; but, some of these factors restrict the use of alternative database models. Consequently, certain limiting factors have led to alternative models of databases that are used in these scenarios. Primarily, motivated by the issue of system scalability, a new generation of databases, known as NoSQL, is gaining strength and space both in information systems. The NoSQL databases emerged in the mid-90s, from a database solution that did not provide an SQL interface. Later, the term came to represent solutions that promoted an alternative to the Relational Model, becoming an abbreviation for Not Only SQL.

The purpose, therefore, of NoSQL solutions is not to replace the Relational Model as a whole, but only in cases in which there is a need for scalability and bigdata. In the recent years, a variety of NoSQL databases has been developed mainly by practitioners looking to fit their specific requirements regarding scalability performance, maintenance and feature-set. Subsequently, there have been various approaches to classify NoSQL databases, each with different categories and subcategories, such as key-value stores, column-oriented and graph databases, oriented-document. MongoDB [2], Neo4j [3], Cassandra [4] and HBase [5] are examples of NoSQL databases. This article only applies to NoSQL document-oriented databases, because of the heterogeneous characteristics of the each NoSQL database classification.

Nonetheless, data modeling still has an important role to play in NoSQL environments. The data modeling process [1] involves the creation of a diagram that represents the meaning of the data and the relationship between the data elements. Thus, ”understanding” is a fundamental aspect of data modeling [6], and a pattern for this kind of representation has few contributions for NoSQL databases.

Addressing this issue, this article proposes a standard for NoSQL data modeling. This proposal uses NoSQL document-oriented databases, including geographic data, aiming to introduce modeling techniques that can be used on databases with document features.

The remainder of the paper is organized as follows: Section II presents related works. Section III explores the concepts of modeling for NoSQL databases based on documents, introducing the different types of relationships and associations. Section IV shows the proposal model in the context of NoSQL databases based on documents. Section V presents the study case to validate the proposal model. Finally in Section VI, we present the conclusion of the research and future works.

II. Related Works

Specifically for geographic data, the current literature suggests a standard of data modeling for relational databases, for example: OMT-G, the acronym for Object Modeling Technique for Geographic Applications [7], GMOD, an environment for modeling and design of geographic applications [8], GISER, a data model for geographic information systems [9], GeoOOA, an object-oriented analysis for geographic information systems [10], MODUL-R, a data model for design spatio-temporal databases [11] and OMT EXT, an explicit representation of data that depends on topological relationships and control over data consistency [12].

Katsov [13] presents a study of techniques and patterns for data modeling using the different categories of NoSQL databases. However, the approach is generic and does not define a specific modeling engine to each database.

Arora and Aggarwal [14] propose a data modeling, but restricted to MongoDB document database, describing a UML Diagram Class format to represent the documents.

Similarly, Banker [15] provides some ideas of data modeling, but limited to MongoDB database and always referring to JSON [16] format as a modeling solution.

As one can see, none of these approaches refers to spatial data, and nor do they present a graphical model for use in any NoSQL document-oriented database.

III. Data Modeling for document-oriented database

An important step in database implementation is the data modeling, because it facilitates the understanding of the project through key features that can prevent programming and operation errors. For relational databases, the data modeling uses the Entity-Relationship Model [1]. For NoSQL, it depends on the database category. The focus of this article is NoSQL document-oriented databases, where the data format of these documents can be JSON, BSON, or XML [17].

Basically, the documents are stored in collections. A parallel is made with relational databases, the equivalent for a collection is the record (tuple) and for a document it is the relation (table). Documents can store completely different sets of attributes, and can be mapped directly to a file format that can be easily manipulated by a programming language. However, it is difficult to abstract the modeling of documents for the entity relationship model [6].

A. Modeling Paradigm for document-oriented database

The relational model designed for SQL has some important features such as integrity, consistency, type validation, transactional guarantees, schemes and referential integrity. However, some applications do not need all of these features. The elimination of these resources has an important influence on the performance and scalability of data storage, bringing new meaning to data modeling.

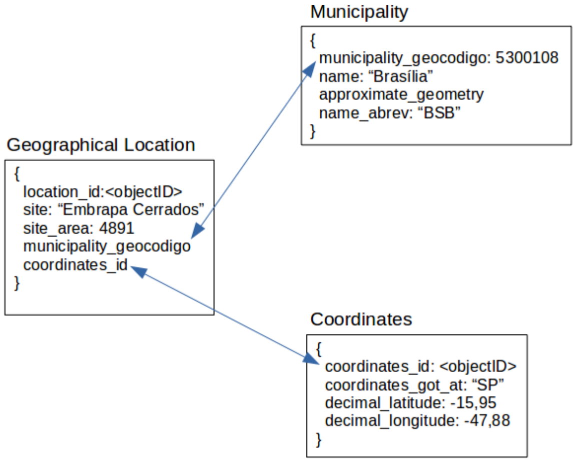
Document-oriented databases have some significant improvements, e.g., index management by the database itself, flexible layouts and advanced indexed search engines [13]. By associating these improvements (some being denormalization and aggregation) to the basic principles of data modeling in NoSQL, it is possible to identify some generic modeling standards associated to document-oriented databases. Analyzing the documentation of the main document-oriented databases, MongoDB [18] and CouchDB [19], similar representations of data mapping relationships can be found: **References** and **Embedded Documents**, featuring a structure which allows associating a document to another, retaining the advantage of specific performance needs and data recovery standards.

B. References Relationship

This type of relationship stores the data by including links or references, literally, from one document to another. Applications can resolve these references to access the related data in the structure of the document itself [18]. Figure 1 shows three documents for Geographical Location, Municipality and Coordinates in a reference relationship.

C. Embedded Documents

This type of relationship stores in a single document structure, where the embedded documents are disposed in a field or an array. These denormalized data models allow data manipulation in a single database transaction [18]. Figure 2 shows a document of a Land Treatment with a Management embedded document.

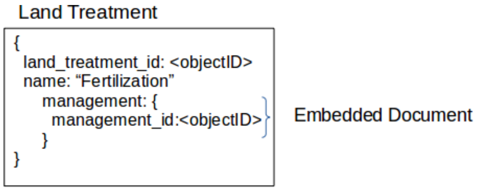


```
{
municipality_geocodigo: 5300108
name: “Brasília”
approximate_geometry
name_abrev: “BSB”
}
```

```
{
location_id:<objectID>
site: “Embrapa Cerrados”
site_area: 4891
municipality_geocodigo
coordinates_id
}
```

```
{
coordinates_id: <objectID>
coordinates_got_at: “SP”
decimal_latitude: -15,95
decimal_longitude: -47,88
}
```

Figure 1. Example of documents referenced



```
{
 land_treatment_id: <objectID>
 name: “Fertilization”
    management: {
      management_id:<objectID>
    }
}
```

Figure 2. Example of embedded documents

IV. Proposal for Document-oriented databases viewing

Unlike the traditional relational databases that have a simple form in the disposition in rows and columns, a document-oriented database stores information in text format, which consists of collections of records organized in the key-value concept, ie, for each value represented a name (or label) is assigned, which describes its meaning. This storage model is known as JSON object, and the objects are composed of multiple name/value pairs for arrays, and other objects.

In this scenario, the number of objects (or documents) in a database increases the abstraction complexity of the logical relationship between the stored information, especially when objects have references to other objects. Currently, there is a lack of solutions to conceptually represent those associated with a NoSQL document-oriented database. As described in [14], there is no standard to represent this kind of object modeling, and several different manners of modeling may arise, depending on each data administrator’s understanding, which makes learning difficult for those who need to read the database model.

Therefore, this section proposes a standard for document-oriented database viewing, including for geographical data references.

Our proposal has some properties, considering the conceptual representation modeling type, such as:

- Ensuring a single way of modeling for the several NoSQL document-oriented databases.
- Simplifying and facilitating the understanding of a document-oriented database through its conceptual model, leveraging the abstraction ability and making the correct decisions about the data storage.

- Providing an accurate, unambiguous and concise pattern, so that database administrators have substantial gains in abstraction, understanding.
- Presenting different types of relationships between documents are defined as References and Embedded documents.
- Assisting the recognition and arrangement of the objects, as well as its features and relationships with other objects.

The following subsections present the concepts and graphing to build a conceptual model for NoSQL document-oriented databases.

A. *Assumptions*

Before starting the discussion about the approach of each type of the conceptual modeling representation, it is important to highlight some basic concepts about objects and relationships in a document-oriented database:

- An object (or document) describes a collection of records that have their properties organized in a key-value structure.
- Information contained in an object is described by the identifier (key) and the value associated with the key.
- Different types of relationships between documents are defined as **References** and **Embedded Documents**.
- Because NoSQL is a non-relational data database, the concepts of normalization, do not apply.
- In contrast, some concepts of relationships between objects are similar to ER modeling, such as cardinality (one-to-one, one-to-many, many-to-many).

B. *Basic Visual Elements*

The proposed solution for a conceptual modeling to the NoSQL document-oriented databases has two basic concepts: *Document* and *Collections*.

As noted previously, a document is usually represented by the structure of a JSON object, and as many fields as needed may be added to the document. For this proposed solution, a document is represented by Figure 3.

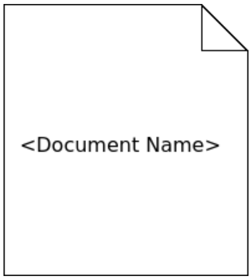


Figure 3. Graphical representation of a Document

It is also possible to store and organize the data as a collection of documents with fields and default values for each record. In this context, a collection of documents will be represented by Figure 4.

The following section presents the definitions of the relationship types and degrees for the objects features.

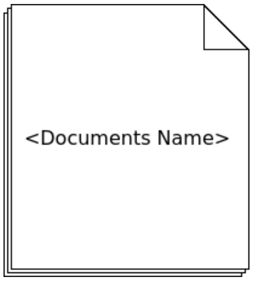


Figure 4. Graphical representation of a Collection

C. *Embedded Documents 1..1*

This section proposes a model that represents the one-to-one relationship for documents embedded in another document. In this case, the proposal is to use the representation of an individual document within another element that represents a Document. In Figure 5, cardinality is also suggested to specify the one-to-one relationship type.

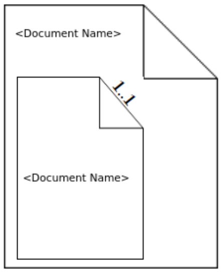


Figure 5. One-to-one relationship for embedded documents

D. *Embedded Documents 1..N*

A one-to-many relationship in embedded documents is represented by the Figure 6. This is the case when the notation to represent the cardinality is the same used in UML [20] and is placed in the upper right corner of the embedded documents. According to the cardinality one-to-many the larger document has embedded multiple documents within it.

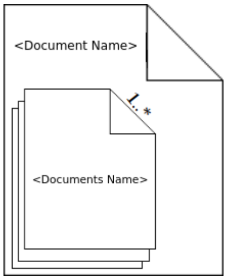


Figure 6. One-to-many relationship for embedded documents

E. *References 1..1*

A document can reference another, and in this case, one must use an arrow directed to the referenced document, as shown in Figure 7. One can see that the directed arrow makes the left document references to the right document. Furthermore, the cardinality of the relationship should be specified above the arrow. The notation of cardinality is based on UML [20].

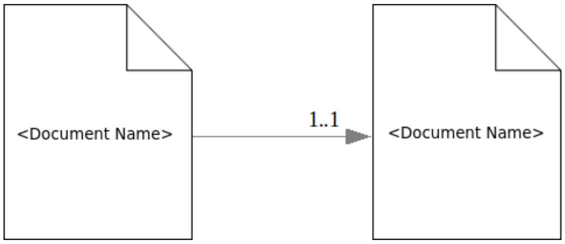


Figure 7. One-to-one relationship for documents referenced

F. References 1..N

In NoSQL, a document can reference multiple documents. To represent this relationship one should use an arrow directed to the referenced documents, as shown in Figure 8. The left document references multiple documents on the right side, by the directed arrow. Furthermore, the cardinality of the relationship is represented by the notation “1..*” as in UML [20].

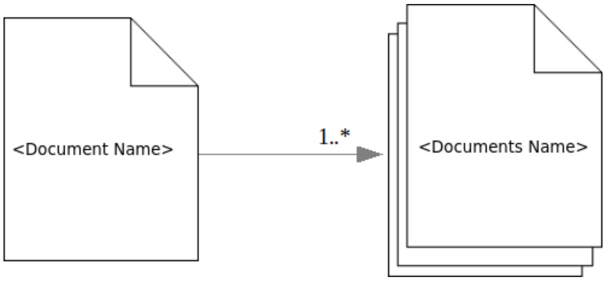


Figure 8. One-to-many relationship for documents referenced

G. Geographical Primitives

With the goal of increasing the capacity of semantic representation of geographic space, the OMT-G [7] model provides primitives to model the geometry of geographic data, so it is possible to georeference documents using these primitives, as shown in Figure 9. A conventional document differs from a georeferenced document precisely by using primitive geographics in the top right corner of the document. In this case, a point is represented by a star, a line is represented by a solid line and a polygon is represented by a square.

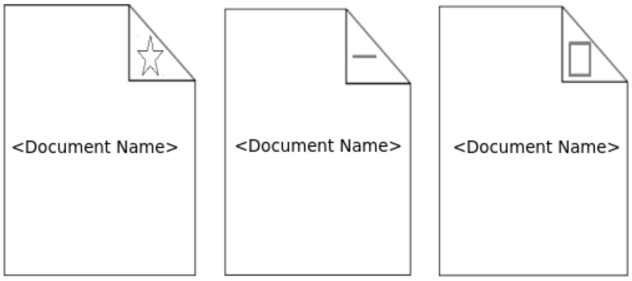


Figure 9. Geographical primitives in documents

Association of the georeference documents is represented by dashed lines [7], as shown in Figure 10. On the other hand, the association of a conventional (not-georeferenced) document with a geo-referenced document is displayed by solid lines, as shown in Figure 11. In both cases, we decided to use the same symbolism for relationship, dashed and solid lines, similar to the OMT-G model [7].

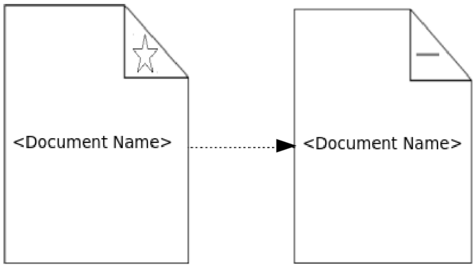


Figure 10. Spatial association

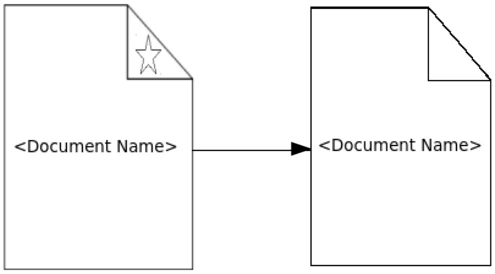


Figure 11. Simples association

Finally, Figure 12 shows the relationship between two embedded georeferenced documents.

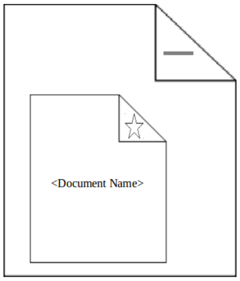


Figure 12. Embedded Document With Georeferences

V. Case Study

An example of data modeling is presented in this section to illustrate the proposed model. This conceptual model is based on the model presented in [21], which corresponds to a collection of information about changes in land use over the Brazilian biomes represented by points (latitude, longitude) where the data was collected.

Thus, the document ”Location” is the document containing georeferenced data such as latitude, longitude and the collection site. A star is used as a geographical point primitive for this document in the upper right corner of the document representing the collection points, as illustrated in Figure 13.

The document “Actual type use” describes the type of real land use (cropland, pasture, secondary forest, etc). While the

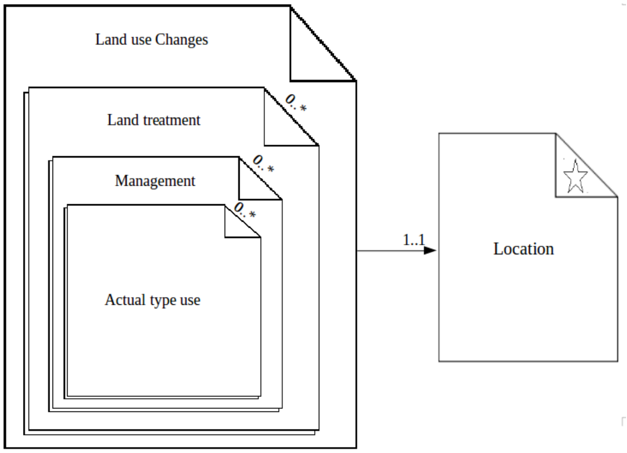


Figure 13. Modeling use case

document “Management” describes the type of land management carried out (burning, fertilizing, planting). The document “Land treatment” describes information relating to the treatment of land (accidental fire, prescribed fire, not burned, etc.). Finally, the document ”Land Use Changes” presents information as species involved, predominant vegetation, removal technique.

Figure 13 shows the relationship between the documents ”management” and ”Actual type use” in which the cardinality is zero-to-many, which means that a document ”Management” may contain zero or more documents of the type ”Actual type use”. Similarly, Figure 13 shows the relationship between the documents ”Land treatment” and ”Management” whose cardinality is also zero-to-many, meaning that a document ”Land treatment” may contain zero or more documents of the type, ”Management”.

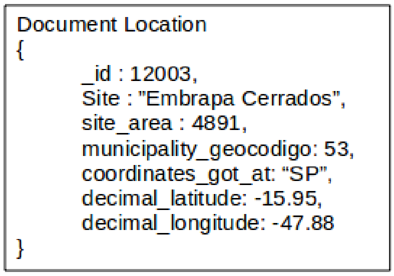
The document ”Land use changes” and ”Land treatment” have a cardinality of zero-to-many, which means that a document of ”Land use changes” may contain zero or more documents of the type, ”Land treatment”.

Finally the document ”Land use changes” reference to the document ”Location” with a cardinality of one to one, means that a document ”Land use changes” can be associated with one location only.

A. Implementation

Considering that the focus of our work is data modeling for NoSQL database, including spatial data features, we have a small range of NoSQL databases to work with. Although some databases provide extensions for geospatial data, as in Cassandra [22], JBoss Infinispan [23] and ArangoDB [24], only Couchbase [25] and MongoDB [2], both documented-oriented databases, offer good documentation for data modeling. The Couchbase spatial data feature is introduced as an experiment and it is not officially provided by the tool yet [26]. And although Couchbase supports many features, we decided to use MongoDB because it provides an excellent documentation and native mechanisms and indexes for geospatial information.

The MongoDB is a document-oriented and open-source database management system [27]. MongoDB uses a BSON format that is a binary representation of a JSON file. In MongoDB, data is represented as pairs of name-value elements. A field-value’s pair consists of the name of the field and its value, and is always separated by a ”:” character. In the Figure 14, the document ”Location” is represented as a BSON’s format document.



```
Document Location
{
        _id : 12003,
        Site : "Embrapa Cerrados",
        site_area : 4891,
        municipality_geocodigo: 53,
        coordinates_got_at: “SP”,
        decimal_latitude: -15.95,
        decimal_longitude: -47.88
}
```

Figure 14. BSON format representation of document Location

The biomes graphic elements and their spatial references were acquired in digital file format, known as Shape. The practical reference of this research entails installing, configuring and creating a schema in MongoDB database, which is responsible for hosting documents containing JSON collections, and storing the spatial data of Brazilian biomes. The MongoDB transactions are structured and based on a JSON format file, and the scripts created for this research follow this pattern. We created database scripts to enter and access data of the biomes location. Furthermore, it was necessary to index the geospatial columns to improve the performance of the queries operations and to provide data for the tools that are responsible for showing data graphically. The operations to insert and return data from a MongoDB document follow a specific syntax, and the result is an object in JSON format. Considering the correct installation and database configuration, and the proper data entry, the Quantum GIS tool [28], associated with add-ons, is used for graphical representation of data. Two levels of QGIS layers were added. One to represent the Brazilian map and the second containing data collections of Brazilian biomes, which were stored in MongoDB format documents. This representation was possible due to the use of components responsible for integrating Quantum GIS and MongoDB.

The graphical representation result by Quantum GIS is shown in Figure 15. The points represent the geographic location of the samples.

VI. Conclusion and Future Work

In contrast to relational database management systems, NoSQL databases are designed to be schema-less and flexible. Therefore, the challenge of this work was to introduce a data modeling standard for NoSQL document-oriented databases, in contrast to the original idea for NoSQL databeses. The objective was to build compact, clear and intuitive diagrams for conceptual data modeling for NoSQL databases. While the current studies propose generic techniques and do not define a specific modeling engine to NoSQL database, our idea was to present a graphical model for any NoSQL document-oriented database. Moreover, while other studies describe techniques based on UML Diagram Class and JSON format, for example, as a modeling solution, we have proposed a new approach

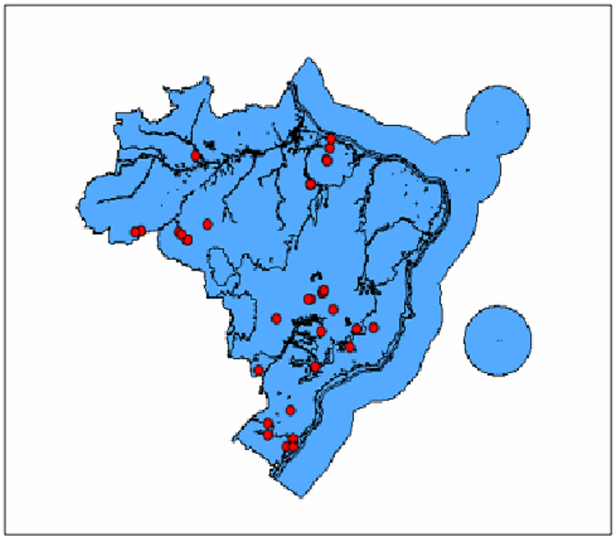


Figure 15. Plotting data in QGIS

to solve the conceptual data modeling issue for NoSQL document-oriented databases, including spatial data references.

Future work includes: verifying our model for other NoSQL database classifications, such as key-value and column; many-to-many relationship for embedded and reference documents is not covered and adjacency, connectivity, and other topological spatial concepts. These points were not developed in the present model proposed in this article.

REFERENCES

[1] R. Elmasri and S. Navathe, Fundamentals of Database Systems. Pearson Addison Wesley, 2010.

[2] MongoDB. Document database. [Online]. Available: http://www.mongodb.org/ [retrieved: Jun., 2014]

[3] J. Partner, A. Vukotic, and N. Watt, Neo4j in Action. O'Reilly Media, 2013.

[4] D. Borthakur et al., "Apache hadoop goes realtime at facebook," in Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011, pp. 1071–1080.

[5] F. Chang et al., "Bigtable: A distributed storage system for structured data," ACM Transactions on Computer Systems (TOCS), vol. 26, no. 2, 2008, p. 4.

[6] R. F. Lans, Introduction to SQL: mastering the relational database language. Addison-Wesley Professional, 2006.

[7] K. A. Borges, C. A. Davis, and A. H. Laender, "Omt-g: an object-oriented data model for geographic applications," GeoInformatica, vol. 5, no. 3, 2001, pp. 221–260.

[8] J. L. De Oliveira, F. Pires, and C. B. Medeiros, "An environment for modeling and design of geographic applications," GeoInformatica, vol. 1, no. 1, 1997, pp. 29–58.

[9] S. Shekhar, M. Coyle, B. Goyal, D.-R. Liu, and S. Sarkar, "Data models in geographic information systems," Communications of the ACM, vol. 40, no. 4, 1997, pp. 103–111.

[10] G. Kosters, B.-U. Pagel, and H.-W. Six, "Gis-application development with geoooa," International Journal of Geographical Information Science, vol. 11, no. 4, 1997, pp. 307–335.

[11] 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, no. 1, 1996, pp. 19–41.

[12] G. Abrantes and R. Carapuca, "Explicit representation of data that depend on topological relationships and control over data consistency," in Fifth European Conference and Exhibition on Geographical Information Systems–EGIS/MARI, vol. 94, no. 19917, 1994, pp. 869–877.

[13] H. Scalable. Nosql data modeling techniques. [Online]. Available: http://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques/ [retrieved: Jun., 2014]

[14] R. Arora and R. R. Aggarwal, "Modeling and querying data in mongodb," International Journal of Scientific and Engineering Research (IJSER 2013), vol. 4, no. 7, Jul. 2013, pp. 141–144.

[15] K. Banker, MongoDB in action. Manning Publications Co., 2011.

[16] D. Crockford, RFC 4627 (Informational) The application json Media Type for JavaScript Object Notation (JSON). IETF (Internet Engineering Task Force), 2006.

[17] S. J. Pramod, "Nosql distilled: A brief guide to the emerging world of polyglot persistence/pramod j. sadalage, martin fowler," 2012.

[18] MongoDB. Data modeling introduction. [Online]. Available: http://docs.mongodb.org/manual/core/data-modeling-introduction/ [retrieved: Jun., 2014]

[19] CouchDB. Modeling entity relationships in couchdb. [Online]. Available: http://wiki.apache.org/couchdb/ [retrieved: Jul., 2014]

[20] G. Booch, J. Rumbaugh, and I. Jacobson, The unified modeling language user guide. Pearson Education India, 2005.

[21] H. V. Olivera and M. Holanda, "A gis web with integration of sheet and soil databases of the brazilian cerrado," in Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on. IEEE, 2012, pp. 1–6.

[22] Cassandra. Cassandra documentation from datastax. [Online]. Available: http://wiki.apache.org/cassandra/GettingStarted [retrieved: Dec., 2014]

[23] J. Infinispan. Get started. [Online]. Available: http://infinispan.org/documentation/ [retrieved: Dec., 2014]

[24] AragonDB. Documentation. [Online]. Available: https://www.arangodb.com/documentation [retrieved: Dec., 2014]

[25] Couchbase. Introduction. [Online]. Available: http://docs.couchbase.com/admin/admin/Couchbase-intro.html [retrieved: Jun., 2014]

[26] ——. Writing geospatial views. [Online]. Available: http://docs.couchbase.com/admin/admin/Views/views-geospatial.html [retrieved: Dec., 2014]

[27] MongoDB. The mongodb 2.6 manual. [Online]. Available: http://docs.mongodb.org/manual/ [retrieved: Jun., 2014]

[28] QGIS. A geographic information system free and open source. [Online]. Available: http://www.qgis.org/ [retrieved: Jun., 2014]

A Segmentation-based Approach for Improving the Accuracy of Polygon Data

Alexey Noskov and Yerach Doytsher
Mapping and Geo-Information Engineering
Technion – Israel Institute of Technology
Haifa, Israel
emails: {noskov, doytsher}@technion.ac.il

Abstract—**The suggested method enables us to improve the accuracy of city planning data by matching it with exact cadastral data. The existing approaches do not work well in the case of partial equality of polygon boundaries. The main idea of the presented algorithm in this paper is based on defining correspondent segments of polygon boundaries and further replacing polygon boundary segments of the non-accurate layer by segments of the accurate data set, segments without pairs are rectified using ground control points. The resulting data contain parts of the accurate data set polygon boundaries, whereas the remaining elements are rectified according to the replaced boundary segments. A review implemented by specialists enables us to say, that the results are satisfactory.**

Keywords-Polyline similarity; geometry matching; shape descriptor; topology.

I. INTRODUCTION

The same objects on a map, which are on an equal scale, could be presented with small differences because of a high diversity of data sources, organizations, users, or software. In an ideal situation, accurate geometries of exiting maps should be used for preparing new data sets or for updating. Usually, in the real world, new maps are digitized without respect to existing data sets. In many cases, data are unavailable, or available with significant restrictions, because of legal, technical, or other reasons. Additionally, even if an accurate data set is freely available, people often do not want to spend time using an existing data set; in most cases they prefer to digitize new geometries on a satellite image or scanned map. These data should be aligned using accurate data sets. This problem is especially sensitive for large-scale maps and plans [7].

Rectifying data using a set of ground control points is a popular way of improving the accuracy of a map [18]. The results of this approach are not satisfactory in many cases, because rectified objects could not be identical to directly measured accurate objects. Another possibility is based on defining correspondent objects on an accurate data set by geometry or attributes and replacing correspondent objects [15]. A serious problem with this approach follows from the fact that objects could be partially similar (e.g., segments of a polygon boundary are same, other parts are different). In contrast to existing approaches, the main idea presented in the paper, an algorithm is based on defining correspondent segments of polygon boundaries and further replacing polygon boundary segments of the non-accurate layer by segments of an accurate data set; segments without pairs are rectified by ground control points. The proposed algorithm could be applied for different polygon datasets with small boundary differences.

The problem is described in the paper using cadastral and city planning maps. A cadastral map is a comprehensive register of the real estate boundaries of a country. Cadastral data are product using quality large-scale surveying with total station, Differential Global Positioning System devices or other surveying systems with centimeter precision. Normally, the precision of maps based on non-survey large-scale data (e.g., satellite images) is lower. City planning data contain proposals for developing urban areas. Most city planning maps are developed by digitizing handmade maps, using space images. Almost all boundaries have small discrepancies in comparison to cadastral maps. It is very important to use exact boundaries or their segments on city planning data from a cadastral map. The approach described in the paper enables us to resolve the problem described.

The developed method consists of several stages: converting polygon layers into topological data format; splitting polylines (polygon boundaries) into segments; defining corresponding maximal segments of polylines; moving segments of land-use boundaries without pairs on cadastral map boundaries and moving centroids of planning data polygons according to surrounding boundaries.

This paper is structured as follows: the related work is considered in Section 2. The source datasets and the process of defining initial variables are described in Section 3. The algorithm of defining correspondent polylines (main part of the approach) is presented in Section 4. The process of compiling of the final map is described in Section 5. The results are discussed in Section 6. The conclusion is presented is Section 7.

II. RELATED WORK

Discrepancy problems on digital maps can be resolved in different ways. Common shape matching techniques are currently used in the raster and vector fields, and sometimes in combination with each other. Several common techniques in the field of Shape Similarity or Pattern Recognition could be applied to the various needs of the matched objects and relevant research questions.

Vector matching techniques can be divided into three main categories.

A. Feature-based matching

This group of methods is based on an object's geometry and shape. The degree of compatibility of objects is determined by their geometry, size, or area. The process is carried out by a structural analysis of a set of objects and comparing whether similar structural analysis of the candidates fits the objects of the other data set [2][13]. In [15], comparison of objects is based on analysis of a contour distribution histogram. A polar coordinates approach for calculating the histogram is used. A method based on the Wasserstein distance was published by Schmitzer et al. [6]. A special shape descriptor for defined correspondent objects on raster images was developed by Ma and Longin [22]. Feature-based matching approaches do not allow for resolving our problem, because they have been developed mainly for single shapes; but, we can use them as part of our approach.

B. *Relational matching*

This group of methods takes objects' relationships into account. In [5], topological and spatial neighborly relations between two data sets, preserved even after running operations such as rotation or scale, were discovered. In relational matching, the comparison of the object is implemented with respect to a neighboring object. We can verify the similarity of two objects by considering neighboring objects. The problem of non-rigid shape recognition is studied by Bronstein et al. [4]; the applicability of diffusion distances within the Gromov-Hausdorff framework [4] and the presence of topological changes have been explored in this paper.

C. *Attributes-based matching*

Matching two data sets' objects by attributes could be very effective if a similar data model is used. Two types of attribute matching could be mentioned: Schema-based [11] and Ontology-based. In [16], an approach based on both types is presented. Attributes-based matching is a specific group of approaches; it can only be applied efficiently in special cases with special data. In most situations it is ineffective.

The merging and fusion of heterogeneous databases has been extensively studied, both spatially [10] and non-spatially [19]. The Map conflation method is based on data fusion algorithms; the aim of the process is to prepare a map which is a combination of two or more maps (often for updating an old map). Map conflation approaches have been presented in [7][12][18].

Computer Vision algorithms are popular in the field of data matching [17]. The Open Computer Vision (OpenCV) framework [3] is widely used today; it provides a number of "out-of-the-box" functions enabling us to detect and compare objects and bindings for popular programming languages (e.g., Python [9]). This makes the OpenCV framework very useful for data-matching tasks. Delaunay Triangulation [14] and Voronoi Polygons [1] are very useful techniques for working with discrete vector data and neighbor analysis. We should also note that in the practice - data is distributed in non-topological formats (e.g., Shape File format) and contains an embarrassment of data analysis, because of a surplus number of objects, duplication of primitives, e.g. polygon boundaries, unexpected gaps between objects etc. We need to use one of the topological data formats presented by Landa [21] to avoid these obstacles. Additionally, two perspective methods could be used in GIS data matching to reduce the time and computer resources required: Genetic Algorithms [20] help to avoid Brute-force operations in some cases; OpenCL technology [8] makes it possible to split a process into a huge number of parallel threads on a video card.

III. DEFINING INITIAL VARIABLES

For implementing and testing our approach, GIS data provided by Survey of Israel have been used. They contain cadastre and land-use city planning polygon shape files covering a part of Harish (a town in the Haifa District of Israel) (Figure 1 depicts source data). Overlaid polygon boundaries of two data sets are presented in Figure 2. From the figure, one can conclude that transformation of lines would not yield positive results, because the gaps are extremely variable - the curved parts of lines consist of different numbers of vertices; thus, even with correct parameters of transformation, the result would not be satisfactory.

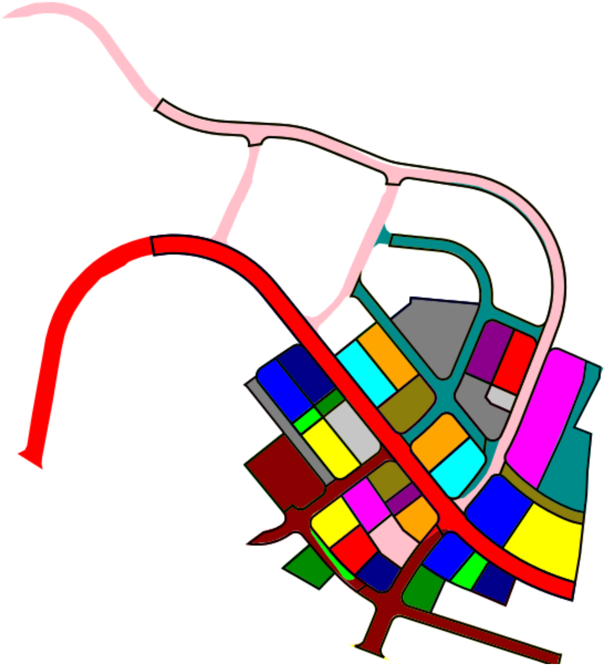


Figure 1. Source data: land-use city planning (color background) and cadastre (black outline) maps.

Source shape files have been converted to GRASS GIS 7 topological data format [21]. Data preparation can be divided into 3 steps:

- Extracting polygon boundaries.
- Splitting polylines into a set of equidistant points. We have decided to use 2 meters between points. For depicting this parameter we will use the symbol d in the paper.
- Calculating an array of distances between the nearest points of two datasets. Setting of initial measures.

Several initial measures need to be calculated. Maximal distance (D_{max}) between the nearest points of two datasets and maximal standard deviation (σ_{max}) have been calculated. To calculate these parameters we need to create a list of 100 percentiles. Then we implement a loop from the first to the last percentile on the list. D_{max} equals percentile i and σ_{max} equals the double standard deviation of distance in the interval between percentiles number i and 100 if the standard deviation of distances between percentiles i-1 and i is more then 1. We calculate tail parameter (t) as follows: t= D_{max}//d. Tail defines a starting or ending segment of polyline which could be ignored.

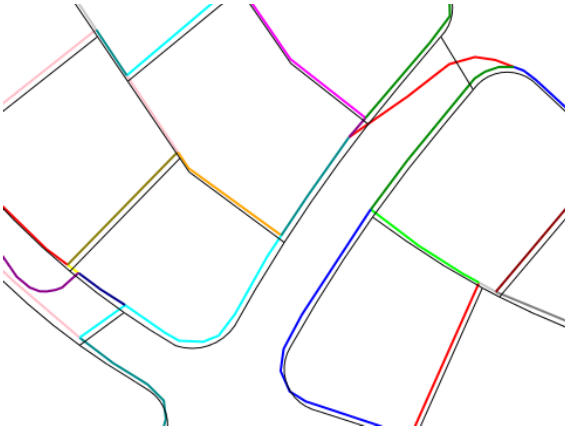


Figure 2. Positional discrepancies of city planning (color lines) and cadastre (black lines) datasets.

We have developed a special shape descriptor (S), partially based on the descriptor presented in (Ma et al., 2011). The descriptor measures the similarity of polylines. Polylines are more similar if S is larger.

$$S=\sum(\exp(-(\log_{10}(1+dists\_a)-\log_{10}(1+dists\_b))^2)+ \exp(-(angles\_a-angles\_b)^2))/matrix\_size^2 \quad (1)$$

In Equation, 1 means matrix of ones, $\log_{10}$ - logarithm with base 10, dists_a – matrix of distances between all pairs of points laid on polyline a. dists_b – matrix of distances between all pairs of points laid on polyline b. If the number of points of a line is k, then matrix size is k×k. angles_a and angles_b are matrices of angles in radians between all pairs of points of lines a and b, correspondingly.

A list containing pairs of point sets has been prepared, where all points laid on line A are closest to points laid on line B of another dataset. For each element of the list, two shape descriptors of tails with t number of points have been calculated and collected into a list of shape descriptors of tails. St_min, St_max – minimal and maximal elements of the list. Also, we use maximal tail standard deviation of point distances (σt), and its (maximal tail) maximal value – σt_max.

The list of initial variables has been calculated: Dmax=10.21, σmax=2.31,t=10.21//2=5, St_min=0, St_max=0.25.

IV. DEFINING CORRESPONDING LINES OF DATASETS

To define corresponding lines, we have developed a special descriptor based on several measures: distances between points, standard deviation of distances, shape descriptor. Figure 3 depicts the main idea – using equidistant points on a polyline to detect corresponding polylines, or segments of polylines. In the figure, a polyline of cadastral data set with nearest polylines of a city planning map are presented.

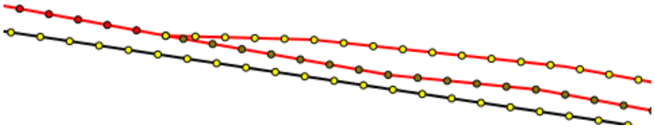


Figure 3. Equidistant points used to calculate similarity of polylines and polylines' segments. Red line – city planning dataset, black – cadastre.

The algorithm for line pairs searching is presented in pseudo code in Figure 4. An explanation follows the listing.

Figure 4. Searching for equal polylines or polylines' segments.

```
Foreach idA,idB in get_ids_of_closest_lines(){
   Pts_A    = get_points('city planning',idA)
   Pts_B =    get_points('cadastre',idB)
   If min(len(Pts_A),len(Pts_B)) > tail {
     Foreach segm in get_segments(Pts_A,Pts_B){
        Pts_A_segm=segm['Pts_A']
        Pts_B_segm=segm[' Pts_B']
        Result_line_pair=find_pair(Pts_A_segm,Pts_B_segm)   } }}
Function find_pair(PtsA,PtsB) {
  If (distance(PtsA[0],PtsB[0]) >
    distance(PtsA[0], PtsB[-1])){ PtsA=reverse(PtsA)   }
  Length=min(len(PtsA), len(PtsB))
  Global_measures=[ ]
  Foreach l in reverse([tail,…,length]){
    Local_measures=[ ]
    Foreach i in [0,…,len(PtsA)-tail]{
      Foreach j in [0,…,len(PtsB)-tail]{
        cur_measure=Calc_measures(PtsA,PtsB,i,j,l)
        if (cur_measure[0]<max_stand_dev and
            cur_measure[1]<max_distance){
          Local_measures.append(cur_measure)} } }
    Global_measures
.append(Find_local_indicator(Local_measures))
    If        Global_measures        and        (l==length        or
len(Global_measures)>tail){
      Gen_desc_list=[calculate_global_indicator(cur) for cur in
                          Global_measures]
      If max(Gen_desc_list[:-tail])> max(Gen_desc_list[-tail:]){
         Return
Global_measures[index_of_maximal(Gen_desc_list)]} }}
Function Calc_measures(PtsA,PtsB,i,j,l){
  cur_PtsA= PtsA[i:i+l]
  cur_PtsB= PtsB[j:j+l]
  dists=Distances(cur_PtsA,cur_PtsB)
  Return [stand_dev(dists),max(dists),
    min(dists),delta_x,delta_y,
   get_max_stddev_of_tailes(cur_PtsA,cur_PtsB),
   get_min_shape_descr_of_tails(cur_PtsA,cur_PtsB),
   get_shape_descriptor(cur_PtsA,cur_PtsB),    i, j, l]}
```

The pseudo-function gets_the 'id's_of_closest_lines () and returns a pair of neighboring lines' ids, points which are closest. Usually, for one line A, several pairs of ids can be defined (idA1-idB1, idA1-idB2,..). All id pairs are processed. Pts_A – points of a city planning dataset are situated on a line with id idA, Pts_B – points of line idB (cadastral map). The pseudo function gets_segments (Pts_A, Pts_B) and splits lines into segments at intervals where the distance between nearest points is more than D_{max}. In the first line of the pseudo function - finding_pairs (PtsA,PtsB) - we test distances from start point of line A to start and end points of line B. If start-start distance is more than start-end, we invert the order of points in line A. Then we set l,i,j variables: l – length of line, i - number of starting point on line A, j - number of starting point on line B. The function Calc_measures (PtsA, PtsB, i, j, l) and calculates a set of parameters (standard deviation of distances, shape descriptor, minimal shape descriptor of line tails, minimal and maximal distance between points). This enables us to define similarity of line A segment from i to i+l and for line B - from j to j+l. Variables i and j which define the optimal segment (pseudo function Find_local_optimal (Local_measures)) have been found for each possible length l using (2).

$$Loc\_Ind=(d-D_{max})/(-D_{max})+(s-St_{\_max})/(St_{\_max}-St_{\_min}) \qquad (2)$$

The meaning of parameters in (2): d – maximal distance between points of lines A and B for (l,i,j), s - minimal tail shape descriptor. In this step we have a Global_measures list containing elements which correspond to some l and contain measures of line segments with maximal indicator Loc_Ind derived from the list (Local_measures) with variable (i,j). This process is illustrated in Figure 5 and 6 (segment length is 20 meters). Figure 5 depicts different segments with the same length; Figure 6 is a plot of indicator Loc_Ind. The next stage is defining optimal segment length. In the previous stage we defined optimal segments i,j for some length l by calculating local indicator Loc_Ind. To define optimal segment length we use global indicator G_Ind; its formula is presented as (3).

$$G\_Ind=((\sigma_t-\sigma_{max})/(-\sigma_{max}))+(d-D_{max})/(-D_{max})+ \qquad (3)$$
$$+(s-St_{\_max})/(St_{\_max}-St_{\_min}))\cdot 2+(1-l)/whole\_segment\_length$$

In the Equation, σ_t means maximal standard deviation of point distances of line segments' tails; for more details see Section 3 and (2). The resulting optimal line length is defined by maximal global indicator G_Ind. The process is illustrated in Figures 7 and 8. Figure 7 depicts examples of optimal segments with different lengths; Figure 8 is a plot of indicator G_Ind. It is obvious that optimal segment length is 41 meters (element with maximal G_Ind, according to plot presented in Figure 8).

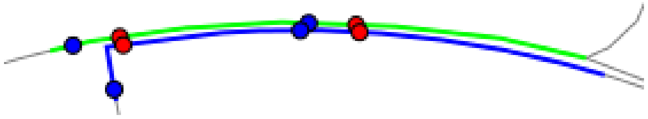


Figure 5. Segment of line A (city planning) – green; segment of line B (cadastre) – blue. Start and end point of the most similar line segments are red points (i=6,j=6, Loc_ind=0.86); blue points – i=2, j=1, Loc_Ind=0.026.

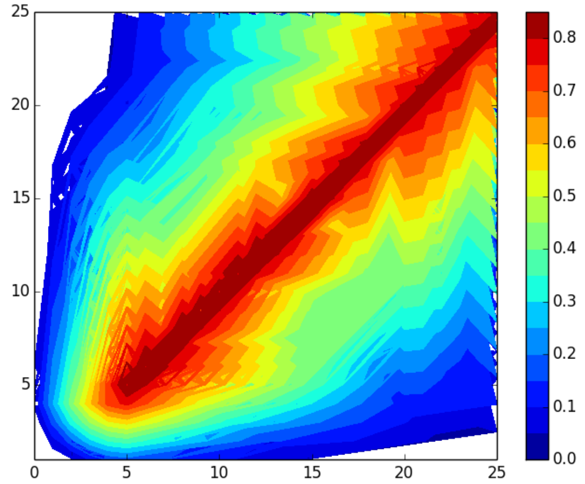


Figure 6. Plot of indicator Loc_Ind: X axis – i, Y axis - j.

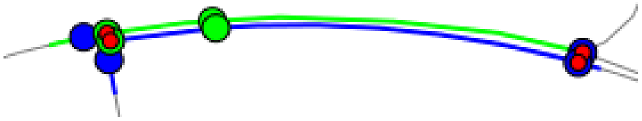


Figure 7. Segment of line A (city planning) – green; segment of line B (cadastre) – blue. Nodes of the most similar line segments with different lengths of segment: red points – l=41,i=5,j=5,G_Ind=2.35; green points – l=10,i=5,j=5 G_Ind=1.69; blue points – l=43,i=3,j=3 G_Ind=1.64.

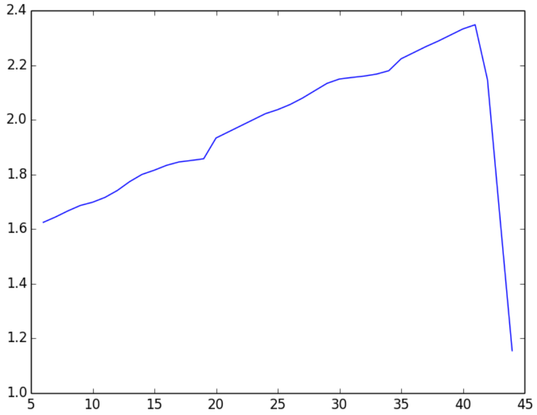


Figure 8. Plot of indicator G_ind: X axis – segment length (in meters), Y axis – G_Ind.

V. COMPILING A FINAL MAP

At this point, we have the pairs of corresponding segments. Some segments are overlapped; to resolve conflicts, a special parameter was developed:

$$P=(l-min\_len)/range\_len+\sigma/(-\sigma_{max}) \qquad (4)$$

where, l is length of line of one of the lines in a line pair, min_len – minimal length of line of all line pairs, range_len – range of length of all line pairs. A line pair with maximal P will be saved, others will be removed. The process is shown in Figure 9.

ISBN: 978-1-61208-383-4

After removing overlapping line pairs, we can use a correspondent line segment of the cadastral dataset instead of the city-planning dataset.

The lines and line segments of the city-planning dataset without the corresponding lines of the cadastral dataset have been moved. Delta X and delta Y have been calculated as average delta X and delta Y of neighboring nodes of line pairs. Unclosed boundaries of polygons have been closed by moving nodes of an unclosed line to the nearest node of a neighboring line (see Figure 10). Centroids of polygons of the city-planning dataset have been moved according to average delta X and delta Y of those boundaries weighted by lengths.

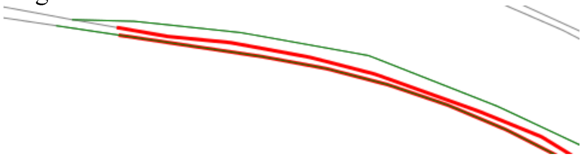


Figure 9. Overlaped line pairs: red line pair – P=1.23, green line pair – P=1.09 . Green line pair will be removed.

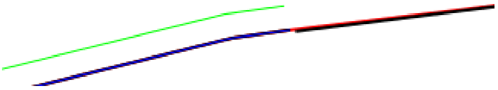


Figure 10. Moving segments without pairs and closing boundaries: green – lines that do not have pairs in cadastral dataset, blue – moved green lines, red – closing boundary by moving nodes, black – cadastral pair of city-planning line segments.

VI. RESULTS

Results are presented in Figure 11 and Figure 12 for two extents. We can conclude that most line segments have been taken from the cadastral dataset; others have been transformed to correspond with cadastral polyline segments. The result looks satisfactory; the final map is holistic and does not contain significant deficiencies. A review implemented by specialists enables us to say, that the results are satisfactory and the approach could be used in real applications after fixing some lacks.

VII. CONCLUSION

An approach for improving the accuracy of polygons' data is presented. The land-use city planning dataset locations have been corrected according to the cadastral dataset. The polylines' segments along the polygons have been split by equidistant points. Analysis has been performed using statistics based on the points of the neighboring polylines of the two datasets. A set of parameters has been used: shape descriptor of polyline segments, standard deviation of point distances, minimal and maximal point distances, standard deviation of segment tails, etc. A set of correspondent polyline segments has been found using special indicators, which enables us to find optimal segments from the list of polyline segments with different length and starting point. The polyline segments of the city planning data with similar/identical parameters to the segments of the cadastral data were linked to these segments (defining counterpart segments). Segments without a counterpart have been transformed.

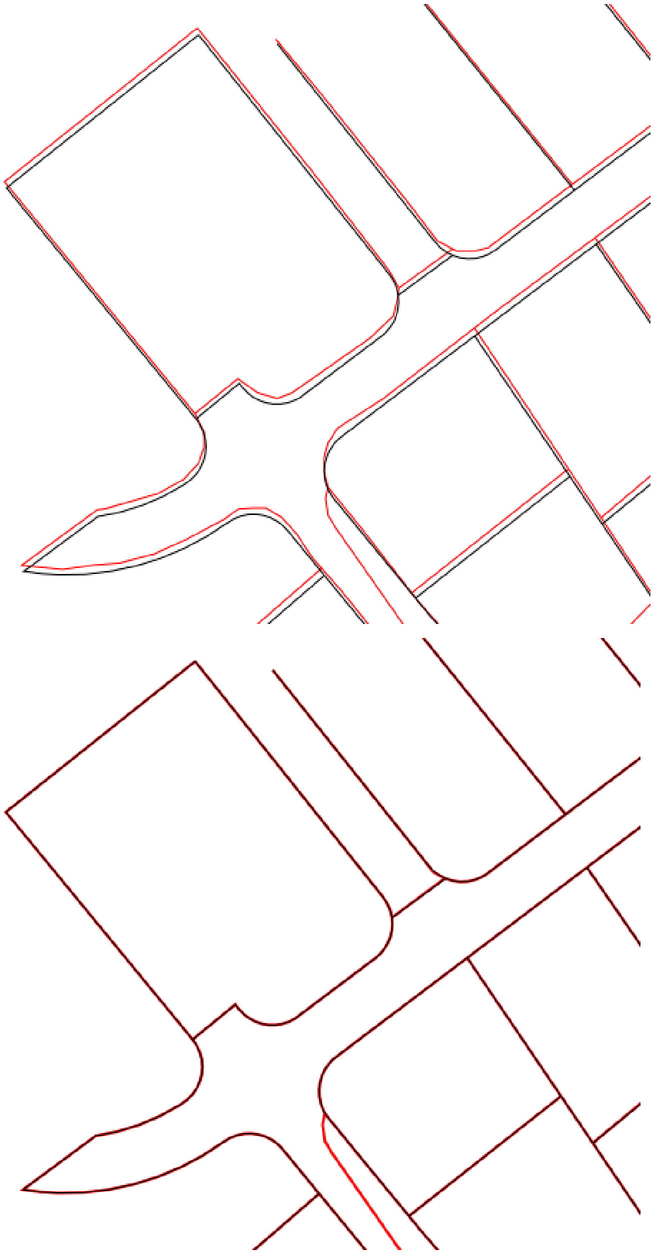


Figure 11. Results. Extent 1. Red – land-use city planning dataset, black – cadastral dataset. Upper – original data, lower –result.

To implement the approach, we used Python 2.7 programming language (with numpy, scipy and matplotlib additional libraries), GRASS GIS 7.1, and Debian GNU/Linux 8 operating system.

In the future, we need to test the approach with more datasets and different parameters, to compare with other approaches, to reduce calculation speed (the execution currently requires about one hour, too long for such a small dataset), and to investigate the reasons for deficiencies and unexpected geometries on the final map.

ACKNOWLEDGEMENT

This research was supported by the Survey of Israel as a part of Project 2019317. The authors would like to thank the Survey of Israel for providing the financial support and data for the purpose of this research.

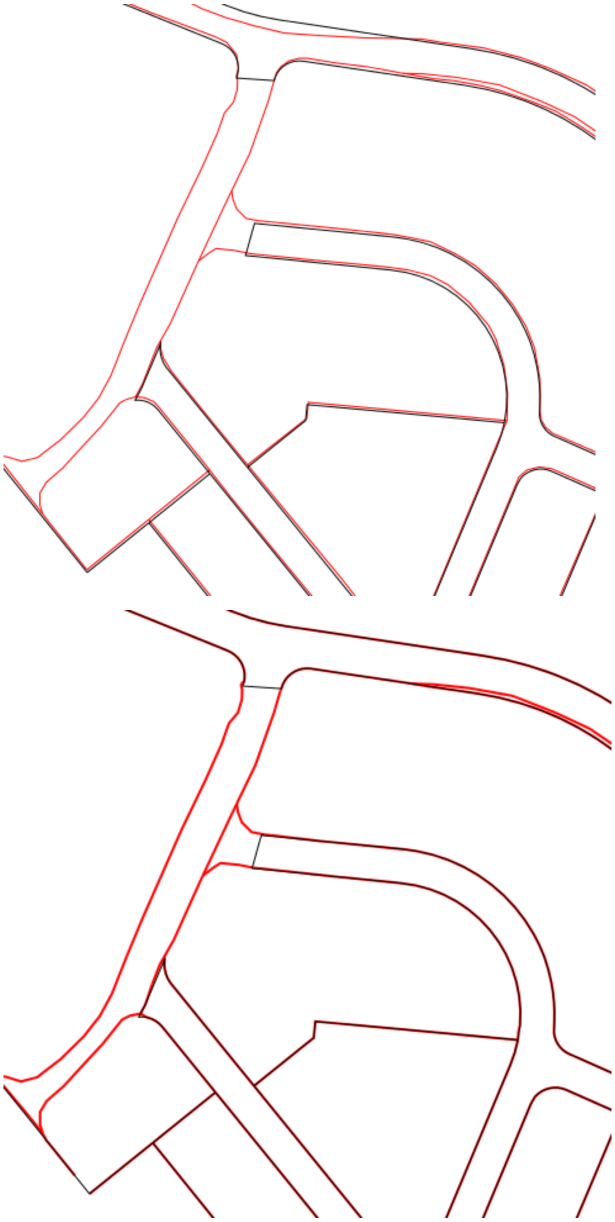


Figure 12. Results. Extent 2. Red – land-use city planning dataset, black – cadastral dataset. Upper – original data, lower –result.

REFERENCES

[1] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," ACM Computing Surveys (CSUR), vol. 23(3), 1991, pp. 345-405.

[2] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(4), 2002, pp. 509—522.

[3] G. Bradski and A. Kaehler, "Learning OpenCV: Computer vision with the OpenCV library," O'Reilly Media, Inc, 2008.

[4] A. Bronstein, R. Kimmel, M. Mahmoudi, and G. Sapiro, "A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching," International Journal of Computer Vision, vol. 89(2-3), 2010, pp. 266-286.

[5] X. Chen, "Spatial relation between uncertain sets," International archives of Photogrammetry and remote sensing, vol. 31(B3), Vienna, 1996, pp. 105-110.

[6] Schmitzer, Bernhard, and S. Christoph, "Object segmentation by shape matching with Wasserstein modes," Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer Berlin Heidelberg, 2013.

[7] S. Filin and Y. Doytsher, "The detection of corresponding objects in a linear-based map conflation," Surveying and land information systems, vol. 60(2), 2000, pp. 117-127.

[8] B. Gaster, L. Howes, D. Kaeli, P. Mistry, and D. Schaa, "Heterogeneous Computing with OpenCL: Revised OpenCL1," Newnes, 2012.

[9] J. Howse, "OpenCV Computer Vision with Python," Packt Publishing Ltd, 2013.

[10] C. Parent and S. Spaccapietra, "Database integration: the key to data interoperability," Advances in Object-Oriented Data Modeling, M. P. Papazoglou, S. Spaccapietra, Z. Tari (Eds.), The MIT Press, 2000.

[11] E. Rahm and P. Bernstein, "A survey of approaches to automatic schema matching," The International Journal on Very Large Data Bases (VLDB), vol. 10(4), 2001, pp. 334–350.

[12] A. Saalfeld, "Conflation-automated map compilation," International Journal of Geographical Information Science (IJGIS), vol. 2 (3), 1988, pp. 217–228.

[13] E. Safra, , Y. Kanza, Y. Sagiv, C. Beeri, and Y. Doytsher, "Ad-hoc matching of vectorial road networks," International Journal of Geographical Information Science, iFirst, 2012, pp. 1–40, ISSN: 1365-8816, ISSN: 1362-3087.

[14] J. Shewchuk, "Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator," Applied computational geometry towards geometric engineering, Springer Berlin Heidelberg,1996, pp. 203-222.

[15] X. Shu and X. Wu. "A novel contour descriptor for 2D shape matching and its application to image retrieval", Image and vision Computing, vol. 29.4, 2011, pp. 286-294.

[16] P. Shvaiko and J. Euzenat, "A survey of schema-based matching approaches," Journal on Data Semantics IV, Springer Berlin Heidelberg, 2005, pp. 146-171.

[17] C. Steger, M. Ulrich, and C. Wiedemann, Machine vision algorithms and applications, Weinheim: wiley-VCH, 2008, pp. 1-2.

[18] V. Walter and D. Fritsch, "Matching spatial data sets: a statistical approach," International Journal of Geographical Information Science (IJGIS), vol. 13 (5), 1999, pp. 445–473.

[19] G. Wiederhold, "Mediation to deal with heterogeneous data sources," Interoperating Geographic Information System, 1999, pp. 1–16.

[20] I. Wilson, J.M. Ware, and J.A. Ware, "A genetic algorithm approach to cartographic map generalisation" Computers in Industry, vol. 52(3), 2003, pp. 291-304.

[21] M. Landa, "GRASS GIS 7.0: Interoperability improvements," GIS Ostrava, Jan. 2013, pp.21-23.

[22] T. Ma and J. Longin, "From partial shape matching through local deformation to robust global shape similarity for object detection," Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. IEEE, 2011, pp. 1441-1448.

Inferring Urban Population Distribution from GSM Data: An Experimental Case Study

Hisham Raslan
Telecommunications Industry Consultant
Teradata Egypt
Giza, Egypt
hisham.raslan@teradata.com

Ahmed Elragal
Business Informatics & Operations Department
German University in Cairo (GUC)
New Cairo, Egypt
ahmed.elragal@guc.edu.eg

***Abstract*-A huge amount of location and tracking data is gathered by tracking and location technologies, such as Global Positioning system (GPS) and Global System for Mobile communication (GSM) devices leading to the collection of large spatiotemporal datasets and to the opportunity of discovering usable knowledge about movement behavior. Movement behavior can be extremely useful in many ways when applied, for example, in the domain of planning metropolitan areas, traffic management, mobile marketing, tourism, etc. Currently, population statistics are derived from census data and household surveys. However, they are difficult to manage and costly to implement. In this paper, we move towards this direction and propose a technique for inferring urban population distribution from GSM data. A case study is used from to build a prototype for testing and evaluating the proposed technique. Results showed that we were able to infer population density with high accuracy based on analyzing GSM data.**

Keywords-Population density; GSM-CDR data; experimental case study.

I. INTRODUCTION

Spatiotemporal patterns that show the cumulative behavior of a population movement can be very useful in understanding mobility-related phenomena. In fact, the discovery of the pattern of traffic flows among sequences of different places in a town can help decision makers to take well informed decisions in different areas, such as urban planning and traffic management. In many application domains, useful information can be extracted from moving object data if the meanings as well as the background information are considered. The knowledge of moving patterns between different places in the geographic space can help the user answer queries about moving objects or movement behavior.

The mainstream literature focuses on finding movement behavior using GPS datasets, however, GPS is rarely used in some developing countries, e.g., Egypt. On the other hand, almost overwhelming majority of the population is using mobile phones (in Egypt, >90M lines which exceed 100% penetration). For this reason, in this research we move towards the direction of using GSM data, as an alternative to the unavailable GPS data.

In this paper, we introduce a technique for identifying home and work locations based on GSM-CDR (Call Detail Records) data. CDR stores the details of the call (caller ID, called ID, time of the call, call duration, etc. Before applying the technique, POI (Points Of Interest) will be identified, and then, the identified POI will be mapped to CDR data, which would then be used in inferring the location. The research relies on experimental case study based on GSM data from a mobile operator to support answering the research question "*how to infer population density from GSM-CDR data*?"

The remaining of this document is organized as follows: Section II provides an overview of related work. Section III describes the proposed framework, conceptual model, and the algorithm used. In Section IV, the proposed framework is applied on an experimental case study. Section V discusses the results of the experimental case study. Section VI evaluates the work done and recommend future research.

II. RELATED WORK

The wide use of technologies, such as Global System for Mobile communication (GSM) networks and Global Positioning System (GPS) resulted in the availability of large amounts of spatiotemporal data. This section provides an overview of related work within the domain of spatiotemporal data analysis and mining research community including related work and research areas that directly address inferring population density using spatiotemporal data.

The identification of dense areas has received the attention. Nevertheless, the techniques used so far suffer from limitations including poor spatial resolution due to the use of grids. Vieira et al. [1] proposed the DAD-MST algorithm to identify dense areas from CDR that is able to process large scale datasets and respects the original tessellation of the space. The DAD-MST algorithm has been tested and validated using a real CDR dataset of almost 50 million entries for over 1 million unique users over a four-month period. The dense areas identified have been qualitatively validated using the subway system of the city under study. The dynamics of the dense areas identified revealed the use that the citizens make of their city, indicating differences between different hour ranges and weekdays and weekends.

One billion people now live in an urban slum, the vast majority of them in developing nations [2]. A slum can be defined as “a residential area which has developed without legal claims to the land and/or permission from the concerned authorities to build; as a result of their illegal or semi-legal status, infrastructure and services are usually inadequate". Sociologists theorize that the majority of urban migration is filtered through slums and understanding the migration patterns is vital to understanding the growth of urban areas. Wesolowski and Eagle [2] used data generated from mobile phones to better understand one of the largest slums, Kibera located in Nairobi, Kenya; focusing on migration patterns out of Kibera, inferring places of work, and tribal affiliations. Kibera has one cell tower location inside the slum, identified with six unique cell tower IDs. In order to form a sample of Kibera's residents, a caller is classified as living in Kibera if they meet all of the following criteria: 1) Over fifty percent of their total calls between the hours of 6 PM and 8 AM have been made from one of Kibera's towers. 2) The total number of calls made in a month is between 3 and two standard deviations from the mean number of calls made by those living in Kibera. They used this identification of callers, looking at three key components of slum dynamics: migration trends, work trends, and tribal affiliations.

Understanding the causes and effects of internal migration is critical to the effective design and implementation of policies that promote human development. Typically government censuses and household surveys do not capture the patterns of temporary and circular migration that are prevalent in developing economies. Blumenstock [3] illustrated how new forms of information and communication technology can be used to better understand the behavior of individuals in developing countries and used mobile phones as a new source of data on internal migration. Using Rwanda as a case study, they developed and formalized the concept of inferred mobility, and computed this and other metrics on a large dataset containing the phone records of 1.5 million Rwandans over four years. The empirical results corroborate the findings of a recent government survey that notes relatively low levels of permanent migration in Rwanda. However, this analysis reveals more subtle patterns that were not detected in the government survey, namely, high levels of temporary and circular migration and significant heterogeneity in mobility within the Rwandan population. The following statistics are computed based on the mobile phone transaction history: Number of cell towers used: As a very crude proxy for the movement of the individual, it is the count of unique towers used by the individual during the specified interval of time. Maximum distance travelled: This is the maximum distance between the set of towers used by the individual over the interval under study. Radius of Gyration (ROG): measures how far an object travels from its center of gravity. In the case of humans, ROG roughly measures the typical range of a user in space.

People spend most of their time at a few key locations, such as home and work. Being able to identify how the movements of people cluster around these "important places" is crucial for a range of technology and policy decisions in areas such as telecommunications and transportation infrastructure deployment. Isaacman et al. [4] proposed a new technique based on clustering and regression for analyzing anonymized cellular network data to identify generally important locations, and to discern semantically meaningful locations such as home and work. The algorithm used for identifying important places has two stages. In the first stage, it spatially clusters the cell towers that appear in a user's trace. In the second stage, it identifies which of the clusters are important using a model derived from a logistic regression of volunteers' CDR's. To select Home or Work, the relevant algorithm (i.e., either the Home or Work algorithm) calculates a score for each important cluster using coefficients obtained from a logistic regression. The algorithm then assigns the cluster with the highest score to be Home or Work.

Understanding the spatiotemporal distribution of people within a city is crucial to many planning applications. Obtaining data to create required knowledge, currently involves costly survey methods [5]. In their research, Toole et al. examined the potential of using GSM-CDR data to measure spatiotemporal changes in population. In the process, they identified the relationship between land use and dynamic population over the course of a typical week. A machine learning classification algorithm was used to identify clusters of locations with similar zoned uses and mobile phone activity patterns. It was shown that the mobile phone data is capable of delivering useful information on actual land use that supplements zoning regulations.

In their attempt to perform a comparative analysis of the behavioral dynamics of rural and urban societies, Eagle et al. [6] estimated the location based on cellular towers. They used four years of mobile phone data from all 1.4M subscribers within a small country. They did not have access to phone numbers, but rather unique IDs that provide no personally identifiable information. In addition to the standard information within CDR, which includes voice and text-messages.

To understand dynamics of human mobility in support of urban planning and transportation management, Phithakkitnukoon et al. [7] developed an activity-aware map that contains most probable activity associated with a specific area in the map based on POIs information. With activity-aware map, they were able to extract individual daily activity patterns from analyzing a large mobile phone data of nearly one million records.

Palmer et al. [8] conducted a pilot study - the Human Mobility Project (HMP) – to explore the use of mobile phones in demographic research and to test a technique: a dynamic, location-based survey. The pilot study uses Global Positioning System (GPS) and cellular tower data from mobile phones to determine subjects' residence locations, observe how they respond to questions at a fixed time and place and also to examine where they spend time when not at home, their trajectories, and how they respond to questions at a variety of different times and places. They were able to recruit 270 volunteers in 13 countries, to share GPS and cellular tower information on their trajectories and respond to dynamic, location-based surveys using an open-source Android application.

The methods currently available for monitoring traffic tend to require the installation of ancillary devices along roadways (loop detectors, cameras and so on), which, along with the costs of installation and maintenance, render alternative techniques more attractive [9]. Their research project has demonstrated a methodology to infer traffic data, such as journey Origin Destination matrices or traffic counts at given points in the road network using a technique that involves analysis of anonymous phone location data in a mobile phone network. The estimation of traffic data has been carried out using non-real data generated by a simulator of vehicular traffic and phones along a stretch of the road network. These simulated data comprise phone location data, which form the source for the development of this technique.

Smoreda et al. [10] reviewed several alternative methods of collecting data from mobile phones for human mobility analysis and described cellular phone network architecture and the location data it can provide. In conclusion, the authors proposed considering cellular network location data as a useful complementary source for human mobility research and provided case studies to illustrate the advantages and disadvantages of each method.

Liao et al. [11] introduced a hierarchical Markov model that can learn and infer a user’s daily movements through an urban community. The model uses multiple levels of abstraction in order to bridge the gap between raw GPS sensor measurements and high level information such as a user’s destination and mode of transportation. Locations, such as bus stops and parking lots, are learned from GPS data logs.

Krumm [12] assessed the privacy threats and countermeasures associated with location data. They examined location data gathered from volunteer subjects to quantify how well four different algorithms can identify the subjects’ home locations and then their identities using freely available, programmable web search engine. Their procedure was able to identify a small fraction of the subjects and a larger fraction of their home addresses. Then, they applied three different obscuration countermeasures designed to foil the privacy attacks: spatial cloaking, noise, and rounding and showed how much obscuration is necessary to maintain the privacy of all the subjects

Advances in sensor networking and location tracking technology enable location-based applications but they also create significant privacy risks. Gruteser et al. [13] proposed a distributed anonymity algorithm that is applied in a sensor network, before service providers gain access to the data. The purpose of these mechanisms is to provide a high degree of privacy, save service users from dealing with service providers’ privacy policies, and reduce the service providers’ requirements for safeguarding private information.

III. Proposed Technique

In our research, we consider population density is a result of residential and business facilities in areas (or POI) and in order to estimate the density we need first to identify the home and work locations for every commuter (user). The question is “how to identify home and work POIs for commuters from the CDRs?”

Every time a mobile subscriber makes or receives a call the network generates a CDR to store the details of the call (caller ID, called ID, time of the call, call duration, etc.). One of the CDR parameters is the cell id, which can be used to identify cell location. It is possible to develop an algorithm to find home and work locations for the commuters based on categorizing day of the week and hour of the day to home, work, travel, and other. Then we aggregate the number of CDRs created by the commuters in different locations at the defined categories; then the location where we find the user most at a specific category, home or work, is considered the commuter home or work POI. Following are the steps to identify commuters home and work POI:

1. Define “location type” as Home, Work, Travel, or Other;
2. “Location Type” value is based on time of the day and day of the week as suggested in Table I;
3. Aggregate the number of CDRs done by commuters for each Location Type per location (POI);
4. For home and work Location Types, the POI with the maximum number of CDRs for a specific location type is assigned as the commuter home or work POI respectively;
5. Knowing home and work POI, we can aggregate to the required level (district, area, city, and governorate).

TABLE I. Definition of location type

Day of Week		Time Interval		Location Type	Time Interval Description
Start	End	Start	End		
1	7	0	6	Home	Home-overnight (Sun - Sat)
6	6	6	11	Home	Home-Friday morning
2	5	20	24	Other	Other-Afternoon Free time
6	6	11	24	Other	Other-Rest of Fri
7	7	0	24	Other	Other-all day, Sat
1	1	0	24	Other	Other-all day, Sun
2	5	6	8	Travel	Travel-Morning rush hour
2	5	16	20	Travel	Travel-Afternoon rush hour
2	5	8	16	Work	Work-regular working hours

In order to apply this technique, we need to identify the mobile operator cells covering the areas under investigation then map their locations to geographical areas (semantics); then, using the CDRs generated with the identified cells over a period of time we can identify operator subscriber’s densities at these areas.

We propose a framework that aims to establish an enhanced approach towards building analysis engine that can infer population density from GSM data. The framework includes process, technologies, data, and decisions as the main aspects towards knowledge extraction.

The phases of the proposed framework are presented in Figure 1.

The process outlines the detailed tasks to take place at each phase and highlights the main layers which data passes through to be transformed into knowledge.

The proposed framework process is comprised of the following phases: storage, data preparation, data pre-processing, analytics, and interpretation. The framework takes all phases into consideration to provide a holistic view of the solution.

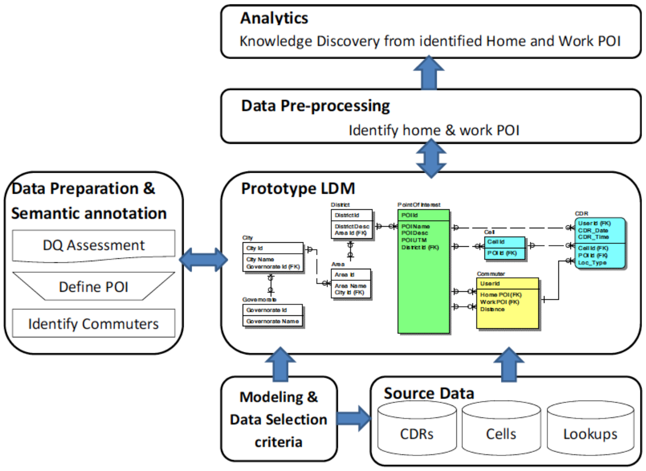


Figure 1. Our Framework.

While using GSM data to study traffic pattern, we will not be violating people's privacy, we will be using disguised data, which do not reveal real identify of people. Besides, unlike GPS data which identify exact location of a subscriber, a GSM cell covers a wide area that makes it impossible to identify the exact location of a subscriber by knowing the location of the cell used during the network activity. This limitation will not affect our research as we are interested in the collective movement behavior between areas.

The following are the tasks performed to reach the results:

1. Build Logical Data Model (LDM);
2. Data loading
   - Load CDRs and Cell information
   - Create and load lookups
3. Data preparation & Semantic annotation
   - Data Quality assessment
   - Define POI to add semantics to the data
   - Identify Commuters (users)
4. Infer Home and Work POI;
5. Perform Analytics.

IV. Experimental Case Study

In this section, we apply the suggested framework on an experimental case study. Our case study is based on data collected, with permission, from a mobile operator in Egypt. The data represents 4 months of CDR's describing Greater Cairo.

A. *Build the Logical Data Model*

The foundation for the required analysis is the Logical Data Model (LDM). The LDM is the conceptual design that shows the business entities and relations between them. It is used to design the Physical Data Model (PDM), which will be created physically on the database management system (DBMS). The LDM is used to help the analytical users plan and develop queries and analytics. Figure 2 shows the LDM used for the case study.

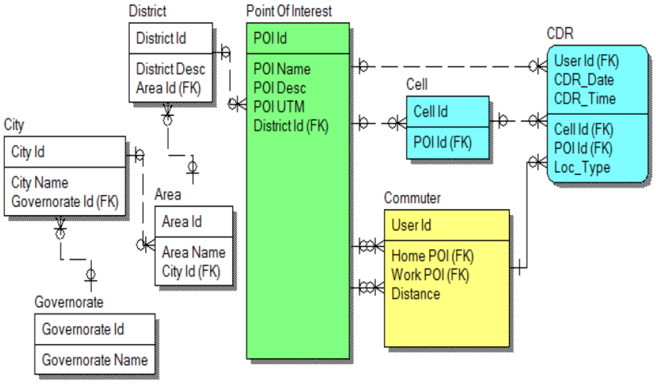


Figure 2. Prototype LDM.

The model is comprised of entities, attributes and relations; the model entities are colored to differentiate the type of information stored in each entity; blue entities are data loaded from the source (CDRs and Cells), white (uncolored) entities have semantics information (lookups: governorate, cities, areas, districts), green represents entities identified for analysis (POIs), and yellow entities contain information extracted from the data (Commuters). The LDM is used to create the PDM, which will be created physically on the selected DBMS.

B. *Data Loading*

Data loading includes two main tasks: (i) loading CDRs and cell information; and (ii) creating and loading lookups. Load CDRs and Cell information is accomplished by the following activities:

- CDR tables for voice and GPRS are merged into the table "CDR". The CDR table has the following structure: (User_Id, CDR_Date, CDR_Time, and Cell_Id)
  - Four Months of voice and data CDR's are merged into the table "CDR".
  - Total number of CDRs: 10,314,009,634 ≈ 1.5B
  - Greater Cairo cells loaded to "Cell" table in the cell id column; the POI column will be mapped later.
  - Number of cells: 14,175
- After loading the CDR table, cell ids are loaded to the "Cell" table in the cell id column; the POI column will be mapped later. POI's were manually identified using Google earth by locating all cells on the map then grouping the cells in a specific area to a POI as we will explain later. The cell table has the following structure: (Cell_Id, POI_Id).

The create and load lookups fulfill the requirements of the data loading phase by creating those lookup tables:

Governorate, City, Area, and District. They are then uploaded with the information required to add semantics to the analysis results.

C. Data preparation

The first step in the data preparation id the data quality assessment. Data quality is the suitability of data to meet business requirements. Because different organizations and applications have different uses and requirements for the data, data quality requirements will also differ. So data do not have to be perfect, but they need to meet business requirements. We will be assessing the data for the following:

- Consistency (Format and Content)
- Completeness
- Uniqueness
- Integrity

We use data profiling tool to perform the required data quality assessments. Table II and Table III show value analysis for the CDR and Cell tables respectively.

TABLE II. VALUE ANALYSIS FOR THE CDR TABLE

Column Name	Count	null	Unique	Zero	Positive	Negative
User_ID	10314009634	0	15424023	0	10314009634	0
CDR_Date	10314009634	0	122			
CDR_Time	10314009634	0	86400	138790	10313870844	0
Cell_ID	10314009634	0	15077	0	10314009634	0

TABLE III. VALUE ANALYSIS FOR THE CELL TABLE

Column Name	Count	null	Unique	Zero	Positive	Negative
Cell_ID	14175	0	14175	0	14175	0
Cell_Status	14175	151	4			
POI_Id	14175	0	522	0	14175	0

In our experimental dataset, CDR's with cell ids that do not exist in the cell table represent:

1. Number of CDRs with missing cells: 214026321 (2% of total CDRs)
2. Number of missing cells: 1127 (8% of total Cells)

The ratios of discovered data issues (2%) will not affect the accuracy of the analysis results.

Second step in the data preparation is to define POI's. This step is comprised of the following:

1. Group cell sites to define POI: POI's under investigation are defined by nearest cell sites. To locate cell sites, all cells were added to Google earth. To add the cells to the map we use the site coordinates. The coordinates are in Universal Transverse Mercator notation (UTM). Cell sites usually have more than one cell and they all have same UTM. The process of grouping cells to a centralized POI is a manual task. We tried to choose the POIs to be close to a real POI and in the same time to identify the urban areas. This process reduced the number of points on the map from around 14K to about 500 points, which is more manageable and easier to locate on the map. The identified POI and the POI demographics are then loaded in the POI table.
2. Update POI column in the cell and CDR tables: As described earlier, the grouping of the cells to POI is done manually and a mapping table was created to map cells to POIs. The mapping table is used to populate the POI column in the cell table and by joining the cells table and the CDR's table on the cell id column we can update POI column in the CDR table.

Third and last step here is to identify commuters. Commuters are the users of the mobile devices who generate voice and data CDR's. To maintain the privacy of the subscribers we only have anonymous identification number (user id) for the subscribers in the CDR table. We can use user id to select the CDRs made by each user (subscriber) and use it for the purpose of this research. To populate the commuter table we select the distinct users from the CDR table and load them in the commuter table. The number of commuters (users) identified is 13,887,256 which is a considerable number of commuters, relative to greater Cairo population, that can fairly represent greater Cairo population distribution.

D. Infer Home and Work POI

Here, the proposed technique described earlier is applied to infer home and work POI in the commuter table. After populating the table, exploration queries are executed to assess the technique results. Resulting indicated that 71% of commuters where identified with home and work POI's (approximately 9.8M); 7% with POI only; 18% with Work only; and 4% with neither. Additionally, and in same process, approximately 10.8M home POI's were identified; 12.3 work POI's. 6.7M commuters were identified with identical home and work. After populating the table, we executed exploration queries to assess the technique results; the resulting figures of the queries are shown in Table IV.

TABLE IV. DATA EXPLORATION RESULTS

Commuters with Home_POI and Work_POI identified	9,831,746	71%
Commuters with Home_POI only identified	950,244	7%
Commuters with Work_POI only identified	2,463,754	18%
Commuters with neither Home_POI nor Work_POI	641,512	5%
TOTAL	13,887,256	100%
Identified Home_POI	10,781,990	78%
Identified Work_POI	12,295,500	89%
Unidentified Home_POI	3,105,266	22%
Unidentified Work_POI	1,591,756	11%
Commuters with identical Home_POI and Work_POI	6,796,530	49%

E. Perform Analytics

By aggregating per the required dimension (POI, District, Area, City, or Governorate) on Home or Work POI, then sorting the results descending, we were able to get the answer for different business question e.g., top busy cities as well as top busy governorates. In Table V, we provide results for Governorates as a sample.

TABLE V. TOP BUSY GOVERNORATES

Home (Governorate)				Work (Governorate)			
Id	**Governorate**	**Total**	**%**	**Id**	**Governorate**	**Total**	**%**
1	**Cairo**	5,361,306	39%	1	**Cairo**	6,444,982	46%
2	**Giza**	3,103,579	22%	2	**Giza**	3,365,887	24%
3	**Qalubia**	2,317,105	17%	3	**Qalubia**	2,484,631	18%
	Unidentified	**3,105,266**	**22%**		**Unidentified**	**1,591,756**	**11%**
	TOTAL	**13,887,256**			**TOTAL**	**13,887,256**	

The numbers are inferred from four month of usage data from June till September 2012. The inferred numbers can be related to the actual numbers as follows:

- The inferred populations of the residence of Cairo, Giza and Qalubia governorates as shown in the table above are about 5M, 3M, and 2M respectively
- According to Egypt census bureau, the population of Cairo, Giza and Qalubia governorates are about 9M, 7M, and 5M respectively [14]
- The percentages for the inferred population relative to the announced are: 55%, 42%, and 40% and these are reasonable percentages for the following considerations:
    - The data used are coming from one operator out of three operators in Egypt;
    - The operator form which we have obtained the data, has been the first to start its operations in Cairo which enabled it to acquire the largest customer base among the three operators;
    - Mobile penetration in Egypt by Q3-2012 was about 113% [15].

As far as movement between cities is concerned, home_POI and work_POI were used to calculate the number of commuters moving between them and aggregating on the dimensions district, area, city, and governorate; we will be able to calculate the volume of traffic between each as follows:

- Select distinct home_POI, work_POI and Count of rows from commuter table to get the number of commuters moving daily from the two POIs
- Sort the results descending on the resulted count will show the two POIs with the highest volume of traffic
- Similarly, do the same to get governorates with the most traffic in and out.

A matrix that shows the movement between governorates is shown in Table VI.

TABLE VI. MOVEMENT BETWEEN GOVERNORATES

Home	**WORK**						
	Giza	**%**	**Cairo**	**%**	**Qalubia**	**%**	**TOTAL Identified Home**
Giza	2,447,841	87%	311,390	11%	43,562	2%	2,802,793
Cairo	214,168	4%	4,532,280	93%	146,419	3%	4,892,867
Qalubia	64,653	3%	261,946	12%	1,809,487	85%	2,136,086
TOTAL Identified Work	2,726,662		5,105,616		1,999,468		9,831,746

V. ANALYSIS OF RESULTS

In this paper, we presented a new technique for inferring population density from spatiotemporal data. The technique could be used in relation to various applications e.g., urban planning and traffic management. We have shown, conceptually and practically, how the technique was able to population distribution in greater Cairo and hence being able to help answer related business questions with regards to movement pattern between cities.

Results showed that it has been possible to infer the population density in areas, cities, and governorates with high accuracy. The analysis can also explain the regular movement volumes of commuters from home to work, which is very useful in identifying movement between areas, cities, or governorates and calculating the average distance of the commuters.

The proposed technique was able to detect movement patterns from high-granularity GSM data whereas mainstream literature focuses on finding these patterns from GPS fine-grained data.

VI. CONCLUSION AND FUTURE WORK

In this research, we proposed a new technique to infer population distribution and movement based on GSM-CDR dataset. While mainstream literature focus on finding this knowledge based on GPS data, we have been able to show how this could be achieved based on GSM data. In the case study, we used real GSM data that was provided by a mobile operator in Egypt. Our dataset included around 4 months of data for 13 million users performed over 10 billion calls and GPRS sessions. We performed different types of analyses covering population density and traffic between areas, cities, governorates. Urban planning and traffic management decision makers could use our technique to make related decision.

Our analysis confirms that long-term GSM activity data is well-suited typical population density analysis, especially when GPS data is not available.

Using real data in the case study was definitely for the benefit of the research. However, the data size was a major obstacle that caused the research to halt several times before Teradata granted the use of one of its servers to the research. This is very important to mention as GSM CDRs are always huge in volume and for deployment the required volume of data will be much more as data from all operators should be integrated to provide complete view for whole population.

Future work includes the deployment of our technique and in different business domains as well as conducting a comparative study between GSM versus GPS data.

REFERENCES

[1] M. R. Vieira, V. Frias-Martinez, N. Oliver and E. Frias-Martinez, "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics," in IEEE Second International Conference, Minneapolis, Aug. 2010, pp. 241-248.

[2] A. P. Wesolowski and N. Eagle, "Parameterizing the Dynamics of Slums," in AAAI Spring Symposium: Artificial Intelligence for Development, Palo Alto, Mar. 2010, pp. 103-108.

[3] J. Blumenstock, "Inferring Patterns of Internal Migration from Mobile Phone Call Records:Evidence from Rwanda," Information Technology and Development, vol. 18 no. 2, 2012, pp. 107-125.

[4] S. Isaacman, R. Becker, R. Caceres and S. Kobourov, "Identifying Important Places in People’s Lives from Cellular Network Data," in Proc. of 9th International Conference on Pervasive Computing (Pervasive), Berlin, 2011, pp. 133-151.

[5] J. L. Toole, M. Ulm, M. C. González and D. Bauer, "Inferring land use from mobile phone activity," in In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, Aug. 2012, pp. 1-8.

[6] N. Eagle, Y. de Montjoye and L. M. Bettencourt, "Community computing: Comparisons between rural and urban societies using mobile phone data," in CSE'09. International Conference, Vol. 4, Vancouver, Aug. 2009, pp. 144-150.

[7] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," in Human Behavior Understanding, Istanbul, 2010, pp. 14-25.

[8] J. R. Palmer, T. J. Espenshade, F. Bartumeus, C. Y. Chung, N. E. Ozgencil and K. Li, "New approaches to human mobility: Using mobile phones for demographic research," Demography, vol. 50, no. 3, 29 Nov. 2013, pp. 1105-1128.

[9] N. Caceres, J. P. Wideberg and F. G. Benitez, "Deriving origin–destination data from a mobile phone network," Intelligent Transport Systems, IET, 1(1), Mar. 2007, pp. 15-26.

[10] Z. Smoreda, A.-M. Olteanu-Raimond and T. Couronné, "Spatiotemporal data from mobile phones for personal mobility assessment," in Transport survey methods: best practice for decision making, London, 2013, pp. 1-20.

[11] L. Liao, D. J. Patterson, D. Fox and H. Kautz, "Learning and inferring transportation routines," Artificial Intelligence, vol. 171, no. 5–6, Apr. 2007, pp. 255-362.

[12] J. Krumm, "Inference attacks on location tracks," in In Pervasive Computing, Toronto, May 2007, pp. 127-143.

[13] M. Gruteser, G. Schelle, A. Jain, R. Han and D. Grunwald, "Privacy-Aware Location Sensor Networks," in HotOS, Vol. 3, Lihue, Hawaii, May 2003, pp. 163-168.

[14] Egypt census bureau, "Publication Name:Population," January 2013. [Online]. Available: http://www.capmas.gov.eg/pdf/EgyptInFigure/EgyptinFigures/Tables/English/pop/population/index.html. [Accessed Jan 2015].

[15] Egypt ICT Indicators, "September 2012 English.pdf," September 2012. [Online]. Available: http://www.egyptictindicators.gov.eg/en/Publications/PublicationsDoc/September%202012%20English.pdf. [Accessed Jan 2015].

Sensors Placement in Urban Environments Using Genetic Algorithms

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

Abstract—**Optimized coverage using multi-sensors is a challenging task, which is becoming more and more complicated in dense and occluded environments such as the urban environments. In this paper, we propose a multi-sensors placement solution for optimized coverage in dense urban environments. Our main contribution is based on a unique concept when facing partially visible objects, such as trees, in an urban scene, extending our previous work and proposing fast and exact 3D visible volumes analysis in urban scenes based on an analytic solution. We consider several 3D models for 3D visibility analysis and present an optimized solution using genetic algorithm, suited to our problem's constraints. We demonstrate the results through simulations with a 3D neighborhood model, taking trees into account.**

Keywords*- Visibility; 3D; Urban environment; Spatial analysis; Genetic algorithm; Sensor coverage.*

I. Introduction and Related Work

Modern cities and urban environments are becoming denser and denser, more heavily populated and are still rapidly growing, including new infrastructures, markets, banks, transportation etc.

In the last two decades, more and more cities and mega-cities have started using multi-camera networks in order to face this challenge, mounting cameras for security monitoring needs; however, this is still not enough [25]. Due to the complexity of working with 3D and the dynamic constraints of the urban terrain, sensors were placed in busy and populated viewpoints, to observe the occurrences in these major points of interest.

These current multi-sensors placement solutions ignore some key factors, such as: visibility analysis in 3D models, which also consist of unique objects such as trees, changing the visibility analysis aspect from visible or invisible states to semi-visible cases, such as trees, and above all, optimization solutions which take these factors into account.

Multi-sensor placement in 3D urban environments is not a simple task. The optimization problem of the optimal configuration of multi sensors for maximal coverage is a well-known Non-deterministic Polynomial-time hard (NP-hard) one [4], even without considering the added complexity of urban environments.

An extensive theoretical research effort has been made over the last four decades, facing a much simpler problem in 2D known as the art gallery problem, with unrealistic assumptions such as unlimited visibility for each agent, while the 3D problem has not received special attention [6].

The coupling between sensors' performances and their environment's constraints is, in general, a complex optimization problem. In this paper, we study the multi-sensors placement optimization problem in 3D urban environments for optimized coverage based on genetic algorithms using novel visibility analysis.

Our optimization solution for this problem relates to maximal coverage from a number of viewpoints, where each 3D position (x, y, z coordinates) of the viewpoint is set as part of the optimized solution. The search space contains local minima and is highly non-linear. The Genetic Algorithms are global search methods, which are well-suited for such tasks. The optimization process is based on randomly generating an initial population of possible solutions (called chromosomes) and, by improving these solutions over a series of generations.

Multi-sensor placements are scene- and application-dependent, and for this reason generic rules are not very efficient at meeting these challenges. Our approach is based on a flexible and efficient analysis that can deal with this complexity.

The total number of sensors is a crucial parameter, due to the real-time outcome data that should be monitored and tracked, where too many sensors are not an efficient solution. We address the sensor numbers that should be set as the tradeoff of coverage area and logical data sources that can be monitored and tracked.

Online visibility analysis is a very complicated task. Recently, off-line visibility analysis, based on preprocessing, was introduced. Cohen-Or et al. [3] used a ray-shooting sample to identify occluded parts.

Since visibility analysis in 3D urban environments is a very complicated task, it is therefore our main optimization function, known as Fitness. We introduce an extended visibility aspect for the common method of Boolean visibility values, "1" for objects seen and "0" for objects unseen from a specific viewpoint, and treat trees as semi-visibility values (such as partially seen, "0.5" value), thereby including in our analysis the real environmental phenomena which are commonly omitted.

We extend our previous work and propose fast and exact 3D visible volumes analysis in urban scenes based on an analytic solution, integrating trees into our 3D model and it is demonstrated with real urban scene model from Neve-Sha'anan (within the city of Haifa) neighborhood.

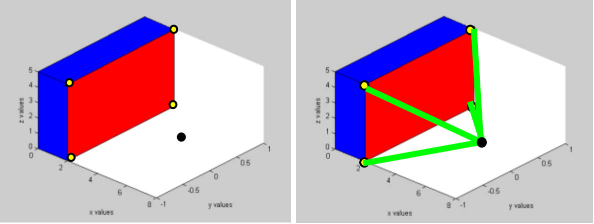
In the following sections, we extended the 3D visible volumes analysis which, for the first time, takes trees into account. Later on, we present the simulation using the Neve-Sha'anan (within the city of Haifa) neighborhood 3D model. Eventually, we present our genetic algorithm optimization stages and simulation based on our 3D visible volumes analysis, taking trees into account.

II. 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 shortly present our analysis presented in [22], extending our previous work [20] 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 [20]. For each object we define Visible Boundary Points and Visible Pyramid. We analyze each building, computing visible surfaces and defining visible pyramids using analytic computation for visibility boundaries [20]. 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 1(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 1. (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 (source: [22]).

In this section, we shortly 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 1(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 2.

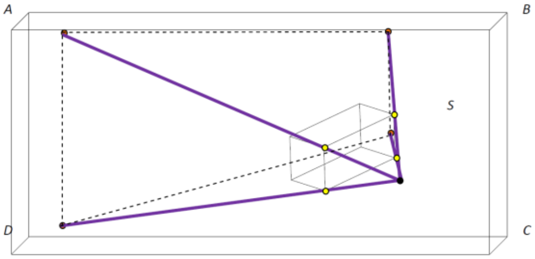


Figure 2. Invisible Projected Visible Pyramid Boundaries colored with purple arrows from a Viewpoint (marked as a Black Dot) to the boundary surface ABCD of Bounding Volume *S* (source: [22]).

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

$$VV_S = V_S - \sum_{i=1}^{N_{obj}} \sum_{j=1}^{N_{surf}} IV_{S_i}^j \tag{1}$$

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

By decreasing the invisible volumes from the total bounding volume, only the visible volumes are computed, as seen in Figure 3. Volumes of VPV and VP can be simply computed based on a simple pyramid volume geometric formula.

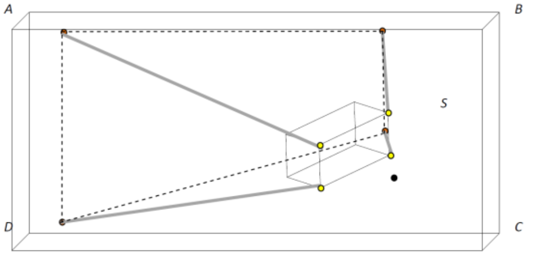


Figure 3. Invisible Volume $V(PVP_i^j) - V(VP_i^j)$ Colored in Gray Arrows. Decreasing Projected Visible Pyramid boundary surface ABCD of Bounding Volume S from Visible Pyramid (source: [22]).

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

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:

$$VV_S = V_S - \sum_{i=1}^{N_{obj}} \sum_{j=1}^{N_{surf}} (V(PVP_i^j) - V(VP_i^j) + IHV_i^j) \qquad (2)$$

The same analysis holds true for multiple overlapping objects, adding the IHV between each two consecutive objects, as can be seen in Figure 4.

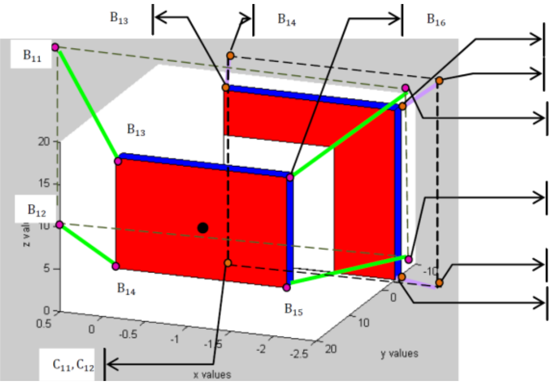


Figure 4. Invisible Volume $V(PVP_i^j) - V(VP_i^j)$ colored in purple and green arrows for each building. PVP of the object close to viewpoint colored in black, colored with pink circles and the far object PVP colored with orange circle (source: [22]).

Extended formulation for two buildings with or without overlap can be seen in [22].

A. Partial Visibility Concept - Trees

In this research, we analyze trees as constant objects in the scene, and formulate partial visibility concept. In our previous work, we tested trees as dynamic objects and their effect on visibility analysis [21]. Still, the analysis focused on trees' branches over time, setting visible and invisible values for each state, taking into account probabilistic modeling in time.

We model trees as two boxes [24], as seen in Figure 5. The lower box, bounded between $[0, h_1]$ models the tree's breed, leads to invisible volume and is analyzed as presented previously for a box modeling building's structures. On the other hand, the upper box bounded between $[h_1, h_2]$ is defined as partially visible, since a tree's leaves and the wind's effect are hard to predict and continuously change over time. Due to these inaccuracies, we set the projected surfaces and the Projected Visible Pyramid of this box as half visible volume.

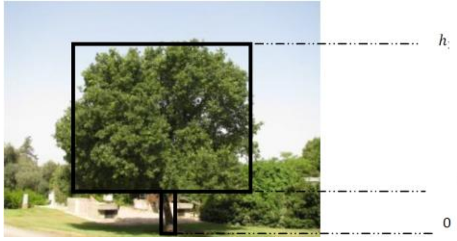


Figure 5. Modeling a Tree Using Two Bounding Boxes.

According to that, a tree's effect on our visibility analysis is divided into regular boxes included in the total number of objects, N_{obj} (identical to the building case), and to the upper boxes modeling tree's leaves, denoted as N_{trees}. The total 3D visible volumes can be formulated as:

$$VV_S = V_S - \sum_{i=1}^{N_{obj}} \sum_{j=1}^{N_{surf}} (V(PVP_i^j) - V(VP_i^j) + IHV_i^j) - \sum_{i=1}^{N_{trees}} \sum_{j=1}^{N_{surf}} \frac{1}{2} (V(PVP_i^j) - V(VP_i^j) + IHV_i^j) \qquad (3)$$

B. Simulations

In this section, we demonstrate our 3D visible volumes analysis in urban scenes integrated with trees, presented in the previous section. We have implemented the presented algorithm and tested some urban environments on a 1.8GHz Intel Core CPU with Matlab. The Neve-Sha'anan Street in the city of Haifa was chosen as a case study, presented in Figure 6.

Figure 6. Views of Neve-Sha'anan Street, Haifa, Israel from Google Maps source: [15]

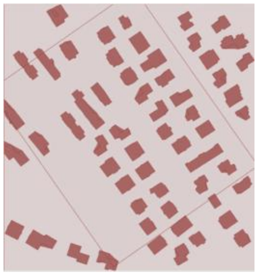


Figure 7. AutoCAD model of Neve-Sha'anan Street, Haifa, Israel.

We modeled the urban environment into structures using AutoCAD model, as seen in Figure 7, with bounding box S.

By using the Matlab©MathWorks software we automated the transformation of data from AutoCAD structure to our model's internal data structure.

Our simulations focused on two cases: (1) small-scale housing in dense environments; (2) Multi-story buildings in an open area. These two different cases are not taking into account the same objects. The first viewpoint is marked with black dot and the second one marked in purple, as seen in Figure 8. Since trees are not a part of our urban scene model, trees are simulated based on similar urban terrain in Neve-Sha'anan. We simulated fifty tree's locations using standard Gauss normal distribution, where trees' parameters $\mathrm{h}_1, \mathrm{h}_2$ are defined randomly $\mathrm{h}_1 \in (0.3,0.9), \mathrm{h}_2 \in (1.5,3)$, as seen in Figure 8.

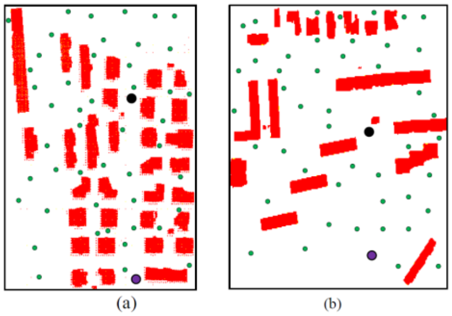


Figure 8. Tested Scenes with Trees marked with green points, Viewpoint 1 Colored in Black, Viewpoint 2 Colored in Purple : (a) Small-scale housing in dense environments; (b) Multi-story buildings in an open area.

We set two different viewpoints, and calculated the visible volumes based on our analysis presented in the previous sub-section. Visible volumes with time computation for different cases of bounding boxes' test scenes are presented in Table II and Table III.

One can notice that the visible volumes become smaller in the dense environments described in Table II, as we enlarge the bounding box. Since we take into account more buildings and trees, less volumes are visible and the total visible volumes from the same viewpoint are smaller.

III. Optimized Coverage Using Genetic Algorithms

The Genetic Algorithm (GA) presented by Holland [23] is one of the most common algorithms from the evolutionary algorithms class used for complex optimization problems in different fields, such as: pharmaceutical design, financial forecasting, tracking and coverage and bridge design. These kinds of algorithms, inspired by natural selection and genetics, are sometimes criticized for their lack of theoretical background due to the fact that in some cases the outcome is unpredictable or difficult to verify.

The main idea behind GA is based on repeated evaluation of individuals (which are part of a candidate solution) using an objective function over a series of generations. These series are improved over generations in order to achieve an optimal solution. In the next paragraphs, we present the genetic algorithms' main stages, adapted to our specific problem.

The major stages in the GA process (evaluation, selection, and reproduction) are repeated either for a fixed number of generations, or until no further improvement is noted. The common range is about 50-200 generations, where fitness function values improve monotonically [23].

TABLE I. Visible Volumes and Computation Time for Small-Scale Housing Case

Bounding Box	Viewpoint	Visible Volumes $[10^5 \cdot m^3]$	Computation Time [sec]
[100 m *100 m * 100 m]	Viewpoint 1	321.7	19.6
	Viewpoint 2	486.8	
[200 m * 200 m * 200 m]	Viewpoint 1	547.4	20.8
	Viewpoint 2	584.2	

TABLE II. Visible Volumes for Small Multi-Story Buildings Case

Bounding Box [100 m *100 m * 100 m]	Visible Volumes $[10^5 \cdot m^3]$	Computation Time [sec]
Viewpoint 1	3453	22.9
Viewpoint 2	3528	

Population Initialization: The initialization stage creates the first generation of candidate solutions, also called chromosomes. A population of candidate solutions is generated by a random possible solution from the solution space. The number of individuals in the population is dependent on the size of the problem and also on computational capabilities and limitations. In our case it is defined as 500 chromosomes, due to the fact that 3D visible volumes must be computed for each candidate.

For our case, the initialized population of viewpoints configuration is set randomly, and would probably be a poor solution due to its random nature, as can be estimated. The chromosome is a 3xN-dimensional vector for N sensor's locations, i.e. viewpoints, where position and translation is a 3-dimensional (x,y,z) vector for each viewpoint location.

Evaluation: The key factor of genetic algorithm relates to individual evaluation which is based on a score for each chromosome, known as Fitness function. This stage is the most time-consuming in our optimization, since we evaluate all individuals in each generation. It should be noticed that each chromosome score leads to 3D visible volume computation N times. As a tradeoff between the covered area and computational effort, we set N to eight. In the worst case, one generation evaluation demands visibility analysis for four thousand different viewpoints. In such a case, one can easily understand the major drawback of the GA method in relation to computational effort. Nevertheless, parallel computation has made a significant breakthrough over the last two decades; GA and other optimization methods based on independent evaluation of each chromosome can nearly be computed in linear time.

Fitness Function The fitness function evaluates each chromosome using optimization function, finding a global minimum value which allows us to compare chromosomes in relation to each other.

In our case, we evaluate each chromosome's quality using 3D visible volumes normalized to the bounding box S around a viewpoint:

$$f(i) = \frac{1}{S}\sum_{j=1}^{N} VV_S(x_j\,, y_j, z_j) \tag{4}$$

Selection: Once the population is sorted by fitness, chromosomes' population with greater values will have a better chance of being selected for the next reproduction stage. Over the last years, many selection operators have been proposed, such as the Stochastic Universal Sampling and Tournament selection. We used the most common Tournament, where k individuals are chosen randomly, and the best performance from this group is selected. The selection operator is repeated until a sufficient number of parents are chosen to form a child generation.

Reproduction: In this stage, the parent individuals chosen in the previous step are combined to create the next generation. Many types of reproduction have been presented over the years, such as crossover, mutation and elitism.

Crossover takes parts from two parents and splices them to form two offspring. Mutation modifies the parameters of a randomly selected chromosome from within a single parent. Elitism takes the fittest parents from the previous generation and replicates them into the new generation. Finally, individuals not selected as parents are replaced with new, random offspring.

A. *Simulations*

In this section, we report on simulation runs with our 3D visible volumes analysis in urban scenes integrated with trees, using genetic algorithms. The genetic algorithms were tested on a 1.8GHz Intel Core CPU with Matlab. We used Fallvile Island Sketchup Google Model [14] for simulating a dense urban scene with trees, as seen in Figure 9.

The stages of Crossover and Elitism operators are described as follows, with a probability of $\mathrm{p_c} = 0.9$ (otherwise parents are copied without change):

1. Choose a random point on the two parents.
2. Split parents at this crossover point.
3. Create next generation chromosomes by exchanging tails.

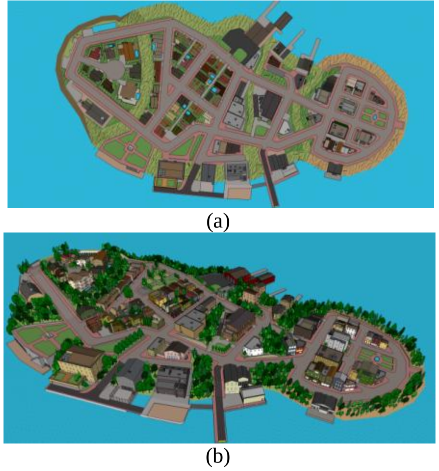
Where the Mutation operator modifies each gene independently with a probability of $\mathrm{p_m} = 0.1$.

In order to process the huge amount of data, we bounded a specific region which includes trees and buildings, as seen in Figure 10. We imported the chosen region to Matlab and modeled the objects by boxes, neglecting roofs' profiles. Time computation for one generation was one hour long on average. As we could expect, the evaluation stage took up 94% of the total simulation time. We set the bounding box S as [500 m* 200 m* 50 m]. Population initialization included 500 chromosomes, each of which is a 24-dimensional vector consisting of position and translation, where all of them were generated randomly.

Based on the Fitness function described previously and the different GA stages and 3D visible volumes analysis, the location of eight viewpoints for sensor placement was optimized. Viewpoints must be bounded in S and should not penetrate buildings and trees. Stop criteria was set to 50 generations and Fitness function gradient.

Optimal coverage of viewpoints and visible volumes during ten running's simulations is seen in Figure 11, bounded in polygons marked with arrows. During these ten running simulations, we initialized the population randomly at different areas inside bounding box S.

These interesting results show that trees' effect inside a dense urban environment was minor, and trees around the buildings in open spaces set the viewpoint's location. As seen in Figure 11, polygon A and polygon B are both outside the areas blocked by buildings. But they are still located near trees, which affect the visible volumes, and we can predict that the same affect will occur in our real world. On the other hand, polygon C, which is closer to the area blocked by buildings, takes into account the trees in this region, but the major factor are still the buildings.


(a)

(b)

Figure 9. Fallvile Island Sketchup Google Model Simulating Dense Urban Scene with Trees, [14]: (a) Topview; (b) Isometric view.

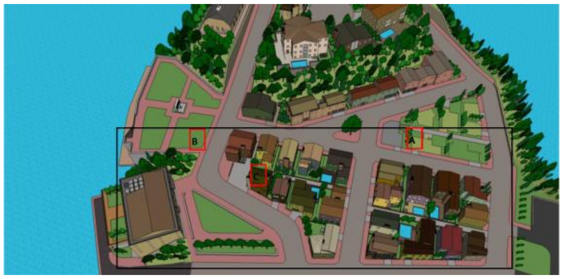


Figure 10. Bounded Area inside Bounding Box S marked in Black, inside Fallvile Island Sketchup Google Model.

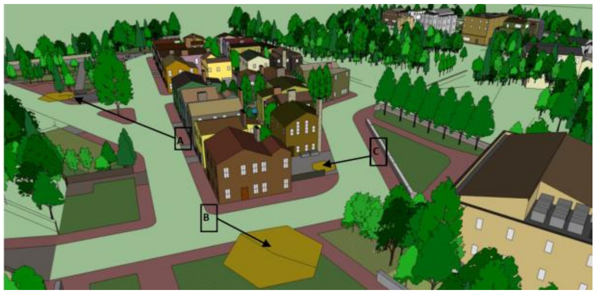


Figure 11. Bounded Polygons of Optimized Cover Viewpoints Using GA marked with Arrows.

IV. CONCLUSIONS

In this paper, we presented an optimized solution for the problem of computing a maximal coverage from a number of viewpoints, using genetic algorithms method. As far as we know, for the first time we integrated trees as partially visible objects participating in a 3D visible volumes analytic analysis. As part of our research we tested several 3D models of 3D urban environments from the visibility viewpoint, choosing the best model from the computational effort and the analytic formulation aspects.

We tested our 3D visible volumes method on real a 3D model from an urban street in the city of Haifa, with time computation and visible volumes parameters.

In the second part of the paper, we introduced a genetic algorithm formulation to calculate an optimized solution for the visibility problem. We used several reproduction operators, which made our optimization robust. We tested our algorithm on the Fallvile Island Sketchup Google Model combined with trees, and analyzed the viewpoint's polygons results.

Our future work is related to validation between our simulated solution and projected volumes from sensors mounted in these viewpoints for optimal coverage.

V. REFERENCES

[1] N. Abu-Akel, "Automatic Building Extraction Using LiDAR Data," PhD Dissertation, Technion, Israel, 2010.

[2] D. Cohen-Or, G. Fibich, D. Halperin and E. Zadicario, "Conservative Visibility and Strong Occlusion for Viewspace Partitioning of Densely Occluded Scenes," In EUROGRAPHICS'98.

[3] D. Cohen-Or and A. Shaked, "Visibility and Dead- Zones in Digital Terrain Maps," Eurographics, vol. 14(3), pp.171-180, 1995.

[4] R. Cole and M. Sharir, "Visibility Problems for Polyhedral Terrain," Journal of Symbolic Computation, vol. 7, pp.11-30, 1989.

[5] Y. Chrysanthou, "Shadow Computation for 3D Interactive and Animation," Ph.D. Dissertation, Department of Computer Science, College University of London, UK, 1996.

[6] R. Church and C. ReVelle, "The Maximal Covering Location Problem," Papers of the Regional Science Association, vol. 32, pp.101-118, 1974.

[7] L. De Floriani and P. Magillo, "Visibility Algorithms on Triangulated Terrain Models," International Journal of Geographic Information Systems, vol. 8, no. 1, pp.13-41, 1994.

[8] L. De Floriani and P. Magillo, "Intervisibility on Terrains," In P.A. Longley, M.F. Goodchild, D.J. Maguire & D.W. Rhind (Eds.), Geographic Information Systems: Principles, Techniques, Management and Applications, pp. 543-556. John Wiley & Sons, 1999.

[9] Y. Doytsher and B. Shmutter, "Digital Elevation Model of Dead Ground," Symposium on Mapping and Geographic Information Systems (ISPRS Commission IV), Athens, Georgia, USA, 1994.

[10] G. Drettakis and E. Fiume, "A Fast Shadow Algorithm for Area Light Sources Using Backprojection," In Computer Graphics (Proceedings of SIGGRAPH '94), pp. 223–230, 1994.

[11] F. Durand, "3D Visibility: Analytical Study and Applications," PhD thesis, Universite Joseph Fourier, Grenoble, France, 1994.

[12] A.E. Eiben and J.E. Smith, "Introduction to Evolutionary Computing Genetic Algorithms," Lecture Notes, 1999.

[13] U. M. Erdem and S. Sclaroff, "Automated camera layout to satisfy task- specific and floor plan-specific coverage requirements," Computer Vision and Image Understanding, vol. 103, no. 3, pp. 156–169, 2006.

[14] Fallvile, (2010) http://sketchup.google.com/3dwarehouse/details?mid=2265cc05839f0e5925ddf6e8265c857c&prevstart=0

[15] D. Fisher-Gewirtzman, A. Shashkov and Y. Doytsher, "Voxel Based Volumetric Visibility Analysis of Urban Environments," Survey Review, DOI: 10.1179/1752270613Y.0000000059, 2013.

[16] W.R. Franklin, "Siting Observers on Terrain," in D. Richardson and P. van Oosterom, eds, Advances in Spatial Data Handling: 10th International Symposium on Spatial Data Handling. Springer-Verlag, pp. 109-120, 2002.

[17] W.R. Franklin and C. Ray, "Higher isn't Necessarily Better: Visibility Algorithms and Experiments," In T. C. Waugh & R. G. Healey (Eds.), Advances in GIS Research: Sixth International Symposium on Spatial Data Handling, pp. 751-770. Taylor & Francis, Edinburgh, 1994.

[18] W.R. Franklin and C. Vogt, "Multiple Observer Siting on Terrain with Intervisibility or Lores Data," in XXth Congress, International Society for Photogrammetry and Remote Sensing. Istanbul, 2004.

[19] H. Furuta, K. Maeda and E. Watanabe, "Application of Genetic Algorithm to Aesthetic Design of Bridge Structures," Computer-Aided Civil and Infrastructure Engineering, vol. 10, no. 6, pp.415–421, 1995.

[20] O. Gal and Y. Doytsher, "Analyzing 3D Complex Urban Environments Using a Unified Visibility Algorithm," International Journal On Advances in Software, ISSN 1942-2628, vol. 5 no.3&4, pp:401-413, 2012.

[21] O. Gal and Y. Doytsher, "Dynamic Objects Effect on Visibility Analysis in 3D Urban Environments," Lecture Notes in Computer Science (LNCS), vol. 7820, pp.147-163, DOI 10.1007/978-3-642-37087-8_11, Springer, 2013.

[22] O. Gal and Y. Doytsher, "Spatial Visibility Clustering Analysis In Urban Environments Based on Pedestrians' Mobility Datasets," The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services, pp. 38-44, 2014.

[23] J. Holland, "Adaptation in Natural and Artificial Systems," University of Michigan Press, 1975.

[24] K. Omasa, F. Hosoi, T.M. Uenishi, Y. Shimizu and Y. Akiyama, "Three-Dimensional Modeling of an Urban Park and Trees by Combined Airborne and Portable On-Ground Scanning LIDAR Remote Sensing," Environ Model Assess vol. 13, pp.473–481, DOI 10.1007/s10666-007-9115-5, 2008.

IQAr: A Web Based GIS for Air Quality Monitoring in Real-time

Matheus Rosendo, Rafael Geha Serta, Yuri Arnold Gruber, Fabiano Scheer Hainosz
Environmental Resources Department
LACTEC Institutes: Curitiba, Brazil
emails: {matheus.rosendo, rafael.serta, yuri.gruber, fabiano.h}@lactec.org.br

***Abstract*— This paper describes the features of the software IQAr that aimed at improving the process of management of air quality data generated by automatic stations. The IQAr includes verification of incoming files for monitoring, processing, validation and storage of data. This tool enables the reporting, the dissemination of results of monitoring over the web with a dynamic map, the visualization of the current direction and speed of wind through a wind rose chart, and the display of the Air Quality Index (AQI) of the last 24 hours. The IQAr allowed the Environmental Institute of Paraná (IAP) and the LACTEC Institutes to improve the equipment evaluation of air quality stations and consistency of the generated data, besides diagnosing air quality in real-time. Due to the ability to continuously validate the data, the software made it possible the dissemination of georeferenced information of air quality in real-time to the general population through the IAP website.**

Keywords-AQI; air quality; real-time monitoring; dynamic thematic map.

I. Introduction

Recent population boom in urban centers caused an increase in industrial production, the number of vehicles on the streets, and in energy consumption, generating an enormous raise in the use of fossil fuels to meet the demand of people. This led to an increase in the level of air pollution and the degradation of air quality as consequence, which has been damaging the health of the population.

Data concerning the monitoring of air pollutant concentrations should be measured in places where the population is exposed to air pollution. Generally, the levels of air pollution are higher in places close to highways (mobile sources) and industries (stationary sources) [1].

The analysis of monitoring data allows evaluating the tendencies of pollutants daily and long term, besides the observation of the effectiveness of the regulations of air pollution and check compliance with standards established by law in order to improve the air quality management [2]. Monitoring should be used to inform the public of critical situations creating emergency programs [3]. The results of monitoring can be related to levels of traffic or industrial activity for urban planning, assisting in environmental licensing [4].

The importance of achieving agility in the process and reliability of the monitoring data of air quality has recently driving interesting and promising research in the area, involving the use of different technologies, such as artificial intelligence for description and classification of air quality [5][6], mobile devices with pollution sensors [7], remote sensing with Geographic Information System (GIS) techniques [8], and wireless sensor networks as an alternative to the conventional measuring stations [9][10].

Automatic monitoring stations of air quality have been playing a key role in the continuous measurement of pollution levels in the atmosphere. However, the amount of data generated by these stations is very large and not always available in a friendly data format for easy interpretation by the end-user.

In order to improve the interpretation, validation and dissemination of information of air quality, it is common for technical experts to use computational resources available in many different types of software, a fact that often makes the process even more costly, time consuming and complex.

Given the above, it is clear the importance to improve monitoring, processing and information management of air quality, as well as make them available to society with simplicity and easy access. The IQAr (acronym of air quality index in Portuguese) aims to meet all these demands.

In this article, the information is organized as follows: The management of data from the air quality stations and the advantages of using IQAr are discussed in Section II. Section III details the developed solution, describes the architecture, each part of the system and its features, including comparisons with the state of the art in air quality. The conclusions, recommendations and future work are described in Section IV.

II. Air quality data management

To manage the raw data received from the air quality stations it is necessary performing mathematical operations in order to generate meaningful information for both expert operator and end-user, like calculating average hourly values for each parameter per station, calculating the Air Quality Index (AQI) to classify the air, converting data in order to compare them to standard of the legislation 03/90 defined by the national environmental council of Brazil (CONAMA), among others.

The air quality standards stipulated by the legislation of each country have the role of protecting the health of citizens, assisting in risk management and environmental policies [1].

The AQI is a dimensionless index used to facilitate the disclosure of pollution data for the general public. This index is also used to get pollutants in the same scale and enable the comparison of the level of pollution between pollutants.

The AQI must be presented together with the referenced period, the name of the critical pollutant, and the category following the indicative color. It also needs to present the concentration of pollutants and their effects on health [11].

The index is calculated accordingly to the concentration and the limits of each region. In Brazil, the air quality index is based on the already mentioned CONAMA 03/90 standard. It is divided into six categories (good, regular, poor, bad, hazardous, critical), where each category represents the level of pollution and health risks. The criteria used follow the primary and secondary standards, level of attention, alert and emergency stipulated in this legislation. The Brazilian system is very similar to the one used in the United States developed by the Environmental Protection Agency (US EPA), where the classification is also divided into six categories: good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy, hazardous [1]. In Canada, the rate is called the index of air quality on health and is used as a tool for the protection of people's health, aiding in decision making to minimize the risk exposure generated by pollution. The index is divided into ten intervals contained in four categories, separated into low, moderate, high and very high risk to health [12].

In 1998, when LACTEC started the operation of the first air quality stations in the state of Paraná, the disclosure was made through an annual report. After the development of the first system by LACTEC for processing and validation of air quality data and partership with environmental institute of Paraná (IAP) it had become possible, in 2010, to release monthly reports.

With the improvement in transmission and validation techniques it had become possible in 2012, the disclosure of daily reports; now, IQAr is in real-time.

In this paper, whenever “real-time” is mentioned, it must be comprehended from 5 to 10 minutes after data are measured by stations, since there is a delay in data transmition by the stations. Regarding air quality area, it should be considered real-time.

Before IQAr, it was not possible for the IAP to provide information about air quality in real-time, since data were not unified into one management system. This required more than one software tool to process, validate and even other software for reporting and disseminating those data. The presented project allowed having all these tasks in one software, making obsolete other systems used previously for this purpose, as Excell, Scanair, Migris and Anagis (previously developed by LACTEC). IQAr joined all the data processed in the various stations of the air quality in a single database. Once incoming data files are at the right format (Figure 2) the processing occurs naturally. Thus, the system is able to process data from stations of different manufacturers, even them using distinct software for measurement and transmission of data.

III. THE IQAR SOFTWARE

The IQAr program was developed by the LACTEC Institutes for IAP in order to manage the periodic measurement data from air quality stations. It analyzes the text files from the stations, processes them, provides features for editing data, performs calculation of the AQI and makes them available on the web using GIS technology in real-time.

A. *System architecture*

As it can be seen in the diagram of Figure 1, data flow begins in the air quality measuring stations.

The stations measure meteorological parameters and pollutants (chemical parameters) in the atmosphere. The pollutants, generally, are: nitrogen oxides (NO_X), sulfur dioxide (SO_2), ozone (O_3), carbon monoxide (CO), total and inhalable particulate matter (MPTS, PM10) and hydrocarbons (HCT). The meteorological parameters are temperature, humidity, wind speed and direction, pressure, global radiation and precipitation.

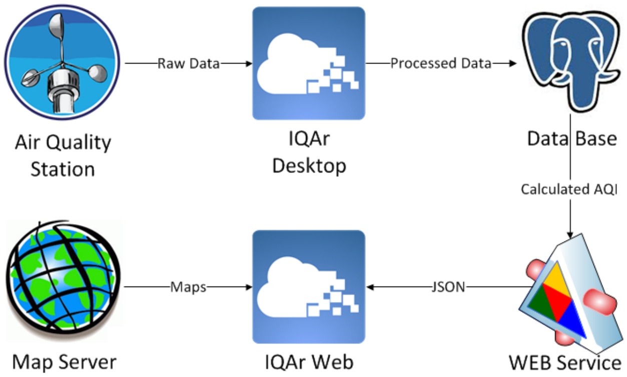
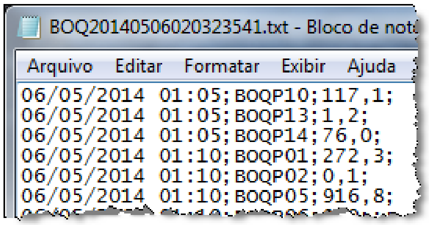


Figure 1. Data flow diagram.

The stations make continuous measurements and store the data in intervals of five to fifteen minutes. Through dial-up line or TCP / IP, they transmit the files to the computer in the data center where it is installed the IQAr Desktop that receives the text files of these stations and performs the processing. In the files, each line represents a measured parameter. The format of these files received by the central consists in date / time, a parameter code and its measured value separated by a semicolon (Figure 2).



```
06/05/2014 01:05;BOQP10;117,1;
06/05/2014 01:05;BOQP13;1,2;
06/05/2014 01:05;BOQP14;76,0;
06/05/2014 01:10;BOQP01;272,3;
06/05/2014 01:10;BOQP02;0,1;
06/05/2014 01:10;BOQP05;916,8;
```

Figure 2. Example of a file generated by a station called BOQ

With the files in the input folder of IQAr Desktop, the data is processed and inserted into the database (more details in Section *C. IQAr Desktop*). Then, the IQAr Desktop calls a stored procedure – a function stored in the database - developed to calculate the hourly average values of each

ISBN: 978-1-61208-383-4

parameter per station and stores this information in another database table, different from the measured raw data table.

This hourly average is a key to the entire system, since the calculations of the AQI for each chemical parameter are based on them. Instead of storing the averages in the bank, the calculation of averages could be done in real-time when the user clicks to generate a report, chart or web viewing; but, in doing so, the time spent to finish the request would be much higher, since the mass of five minutes data (most of the information contained in the database) is 12 times greater than the mass of hourly averages.

Calculated the averages, the data are ready to be accessed by the monitoring module IQAr Web. This access is made via web service that performs the request to the database, calculates the AQI and packages the data in JavaScript Object Notation (JSON) [13] using the main library of IQAr. JSON is an open standard format used to transmit data as an alternative to Extensible Markup Language (XML).

Once the AQI is calculated and returned in JSON format, the monitoring module of IQAr Web updates the information on the map returned from the map server.

Besides, the web service is responsable for all communications between IQAr web and the database, including performing queries to create the wind rose graph, the charts per parameter/station, the monthly reports and the information required by the diagnosis module. This module performs periodical tests to ensure the availability and reliability of the whole system like checking web servers status, database connections, last data inserted, last file processed and last connection done by IQAr Desktop.

Most software developed for air quality management throughout the world has an hourly update routine; which is the case of the South Coast Air Quality Management District system (AQMD) [14]: a web software using GIS to display the state of the air quality in Southern California. On the other hand, the IQAr can provide the information to the end-user within three minutes after being measured and transmitted by air quality stations. This is due to the fact that IQAr manages, fully automatically, from the raw data generated by the stations until the web visualization. For example, nowadays, at most 13 files arrives hourly, one by station. Each transmition takes around one minute and happens just after the hour is completed.

Once the file is ready to be processed, the system takes in average three seconds to process each file which varies according to the number of parameters measured by each station. The hardware used in this processing is a quad core i5 3470 3.2 GH processor with 4GB of RAM memory and Windows 8. IQAr Desktop searches for new data every minute, so the database is always updated within at most one minute and a few seconds of delay in relation to the transmitted data.

The data is then presented to the end-user at web through a JavaScript routine that updates the IQAr Web by default every five minutes, but can be set to do it even more often. For example, setting it up to one minute, the software will display information with less than three minutes of delay since the data arrival from the stations, which is pretty up to date regarding air quality.

It is important to mention that the total delay depends also on the configuration of the measurement frequency and transmitting of each station. For example, if a station sends updated data within 20 minutes, IQAr can ensure that the information displayed was measured within 23 minutes.

B. Used Technology

Most of the system was developed in Java, including the the main library used by the desktop module and web service, as well as the web service and the IQAr Desktop itself. Due to political issues of the project in relation to the data storage, the system had to be designed so that the database and the web application were executed on different servers of different institutions. Thus, we chose to use a web service brokering data between the database and the IQAr Web. This peculiarity made the development more laborious, but, on the other hand, it turned the system into a more portable and flexible solution according to the demand and environmental policy of each institute, since this feature facilitates the adaptation of the system to handle data from other environmental institutes of Brazil.

As database management system, it was used PostgreSQL [15]. To improve the performance of the database and to make it robust enough to hold millions of records it was necessary to use the partitioning of the main tables (raw data and average data). With the partitioning each table holds data for only one month and, as result, every system query became faster than one second. As map server, it was used ESRI ArcGIS Server [16]. The IQAr Web was developed using the ArcGIS API for JavaScript [17] and the PHP language [18]. To transfer information from the web service to the IQAr Web, it used the JSON format. We opted for the JSON instead of XML markup language in order to reduce the amount of information transmitted every request, since the IQAr Web is a monitoring system and makes automated queries often to keep always updated the information on the map. For charts in IQAr Desktop it was used JFreeChart [19] and for charts and other visuals interface effects in IQAr Web it was used the JQuery [20] library.

C. IQAr Desktop

IQAr Desktop is the part of the system responsible for continuously processing data received from stations. It runs on a regular PC within the LACTEC Institutes and is operated by a team of engineers and technical experts in air quality monitoring to ensure the quality of the information processed by the system.

As described in Section III.A, the process starts with the arrival of the station files in the system’s input folder. This directory is checked periodically for new data in a time interval defined by the operator, by default it is set to one minute. When new files are detected, the IQAr proceeds with the format verification of each file, it analyzes if it is empty or truncated, which can occur systematically. In positive case, the IQAr does not process the file and alerts the user in the monitoring log about the detection of an invalid file; otherwise, all lines of the file (Figure 2) are split into information regarding date, time, the measured parameter code and its value in the measurement.

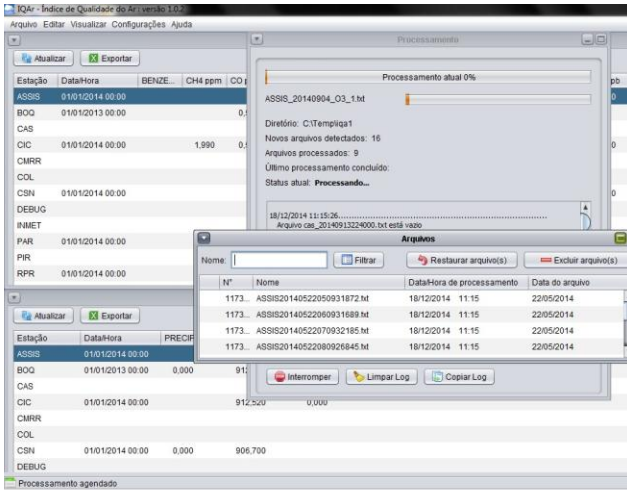
With this information in memory, it begins the process of checking each data contained in the file. Basically, according to predefined chemical and meteorological parameters, three types of validation are made by the system: range, endurance and step. The range validation establishes a minimum and maximum limit for each parameter, for example, in Curitiba, temperature will hardly be less than zero and higher than 35 degrees. The endurance validation verifies if the value is being repeated over a long period, which is generally indicative of a sensor malfunction. Finally, the step validation evaluates large variations over short periods, for example, if the air temperature is close to 20 degrees in a moment, and 30 minutes after, it decreases to 5 degrees, this is an indicative of wrong value; probably, a failure occurred in the station of the air quality. All these validations are adapted to each station because the meteorological and emission conditions are different. The goal is to automatically recognize and classify data errors due to calibration procedures on station, power failure and inconsistent values, allowing data to be disseminated reliably. In the new versions of the IQAr, we plan to implement more types of automatic validation using correlations of chemical data with meteorological data and comparison of data between stations.

In the case of invalid data considered by one of the validations, those data remain available for query and modification by the operator; but, they are not used in the calculation of the indices shown to the end-user.

The system also allows editing of the data processed through screens that allow the operator to perform basic mathematical operations such as addition, subtraction, multiplication and division on a single specific value or all values comprised within a certain period of time. This is important to make manual corrections to the data to ensure the reliability of information. The role of the data validation staff as system operator at this stage is critical; since they are the ones with the expertise to assess whether the data from the stations are correct and after the automatic validation of the system, they can also manually validate or invalidate a block of data from one station. After processing, the files are physically transferred from the incoming directory to the directory of processed files. The system has a settings screen where the operator can set these directories and also the processing interval, which is the time that the system will automatically scan the incoming directory for new data to be processed.

Once the data are stored in the database, the system allows the visualization of information through charts and graphs. In both, the operator can search by station, period and type of data that can be: raw data, hourly average, daily average, hourly AQI or daily AQI. Such information can then be converted into spreadsheets and reports for external use.

As it can be seen in Figure 3, in the system home screen the operator can monitor the latest processed data of all stations divided into two tables: the chemical parameters above and the meteorological parameters below. To monitor the current status of processing, the processing screen displays a status bar with the progress of the file that is being processed; a second bar with the processing status of the complete list of new files found by the system; the amount of files to be processed; the amount of files that have been processed; the date and time of the last completed processing; the current processing status (scheduled, processing, or stopped) and also a log that shows the operator if any problem occurs during inserting data in the database, for example, if there is already a record in the database with the same date, station and parameter.



Estação	Data/Hora	BENZE...	CH4 ppm	CO p
ASSIS	01/01/2014 00:00			
BOQ	01/01/2013 00:00			0,!
CAS				
CIC	01/01/2014 00:00		1,990	0,!
CMRR				
COL				
CSN	01/01/2014 00:00			
DEBUG				
INMET				
PAR	01/01/2014 00:00			
PIR				
RPR	01/01/2014 00:00			

Processamento atual 0%

ASSIS_20140904_O3_1.txt

Diretório: C:\Templiqa1
Novos arquivos detectados: 16
Arquivos processados: 9
Último processamento concluído:
Status atual: **Processando...**

```
18/12/2014 11:15:26.........................................................................
  Arquivo cas_20140913224000.txt está vazio
```

Nome:

Nº	Nome	Data/Hora de processamento	Data do arquivo
1173...	ASSIS20140522050931872.txt	18/12/2014 11:15	22/05/2014
1173...	ASSIS20140522060931689.txt	18/12/2014 11:15	22/05/2014
1173...	ASSIS20140522070932185.txt	18/12/2014 11:15	22/05/2014
1173...	ASSIS20140522080926845.txt	18/12/2014 11:15	22/05/2014

Estação	Data/Hora	PRECIP		
ASSIS	01/01/2014 00:00			
BOQ	01/01/2013 00:00	0,000	91:	
CAS				
CIC	01/01/2014 00:00		912,520	0,000
CMRR				
COL				
CSN	01/01/2014 00:00	0,000	906,700	
DEBUG				

Processamento agendado

Figure 3. Home screen of IQAr Desktop (chemical parameters; meteorological parameters; processed files; current processing).

The operator has total control of all files that have already been processed by the system through the processed files screen. In this screen (also in Figure 3), the operator can search files by name, delete one or multiple files at once and therefore the data from these files. There is also the option to restore after deleting the files, which means removing the file from the processed files folder and put it back in the input directory; this option is useful if the operator wishes to reprocess a file.

IQAr has the ability to work with an unlimited number of stations; it only depends on available hardware resources to process all data. The same applies to the number of parameters per station; there is a screen to control parameters per station where the operator can set it up without limits.

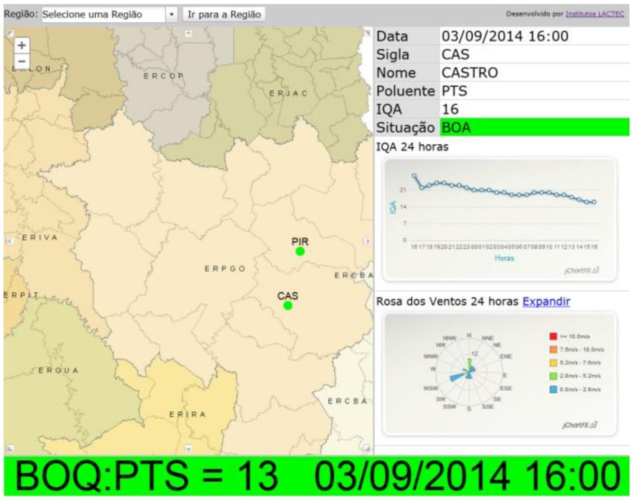
The current air quality stations provide on average 14 parameters, including chemical and meteorological, but this number may be higher depending on setup, demand and technology used in new stations. Previously, the software used to process the data from most stations of the state of Paraná was the Scanair developed by Envitech [21]. The Scanair, unlike IQAr, had a limitation of parameters per station, where no more than 16 parameters could be processed at once by the system.

The system has a configuration screen for each station where the operator can set the geographic coordinates of the station so that the information appears correctly positioned in the map on IQAr Web, and also change parameters that dictate the behavior of the system to calculate the hourly averages of the station. For example, dealing with five minutes data, by default, the system considers valid the hourly averages composed of nine or more valid data on a total of 12, but the operator can configure the system to consider valid hourly averages composed of four valid data in a particular station.

All setup parameters needed to make the average and AQI calculations are customizable by the operator, which makes IQAr flexible and adaptable in case of changes in applicable laws of each country or region. The system can even process data from other systems, such as the case of Scanair itself. A big data exporting was successfully carried out recently: 230 MB of data divided into 247 files exported from Scanair regarding data of nine different stations along 10 years from 2003 to 2013. Even after all those data have been processed it was not noticed any decrease of performance of the system queries, which proves that the partitioning realized on data base (mentioned before) was properly implemented.

D. IQAr Web

The IQAr Web is the part of the system responsible for showing to the end-user the processed information contained in the database.



Região: Selecione uma Região ▾ Ir para a Região

Desenvolvido por Institutos LACTEC

Data	03/09/2014 16:00
Sigla	CAS
Nome	CASTRO
Poluente	PTS
IQA	16
Situação	BOA

IQA 24 horas

Rosa dos Ventos 24 horas Expandir

BOQ:PTS = 13 03/09/2014 16:00

Figure 4. IQAr Web running.

Along with the processed data, geographical coordinates of each station are stored, as shown in Figure 4; they are represented punctually in a dynamic thematic map. Besides the air quality information, the municipal boundaries of the State of Paraná and the regional offices of the IAP are also represented on the map. The condition of air quality for each station is classified according to the following scale and color values:

- Green: good, AQI until 50;
- Yellow: regular, AQI from 51 to 100;
- Orange: inadequate, AQI between 101 and 200;
- Red: bad, AQI between 201 and 300;
- Purple: terrible, AQI between 301 and 400.
- Black: critical, AQI above 400.

This classification was defined by CONAMA and adapted by IAP.

Whenever the user clicks on a station on the map the system shows on the right side of the screen the AQI of the pollutant of highest value for the station in real-time. Since AQI is measured hourly, this value is truncated to the current time, i.e., if the user accesses the system at 15:19, it shows the pollutant with the highest AQI at 15:00. Then, the system displays the name of the pollutant, the corresponding time, the name of the clicked station and the description of the status (good, regular, etc.).

Under the station status, the system shows a chart with the variation of the higher AQI value for the last 24 hours. This way, the responsible by monitoring at the environmental department can have an updated overview of air quality in the city.

It is also presented a wind rose chart indicating the direction and speed of wind for the last 24 hours. This chart allows the monitoring team to identify possible causative of pollution in case of episodes where the air quality is not good and also issue warnings to the population if there is an accident with some toxic gas. At the bottom of the web page, the system shows the current situation of air quality for all stations in the format: "Station: pollutant parameter of higher concentration and its value of AQI, date and time". On the map, the user can zoom with the scroll button or choose a predefined region in the system with several stations contained and use the button "go to the region", so the system automatically positions the map in the given region. The web module is used today to show data from three different institutions, but it can hold data from stations of more institutions from different parts of the country. There is no limitation in terms of software; again depends only on hardware resources to be allocated.

A similar solution to IQAr Web was developed by AQICN [22] in China with the support of the United States Embassy and is being used for dissemination of data from various countries in Asia, Europe, North America and some countries in South America as Brazil, with data being supplied by the company of environmental sanitation technology of São Paulo (CETESB) [23]. The visual quality of the system is very friendly; They used Gmaps [24] as the map server and also developed a version for mobile devices, but the reliability and timeliness of the information shown in the AQICN system are responsibility of the air station owners, and most often environmental institutes do not have software and staff able to provide such information accurately and reliably in real-time. The AQICN software does not replace IQAr in Brazil also because of the standard for calculating the AQI. They use US's EPA index system that is not the one used in Brazil.

IV. CONCLUSION AND FUTURE DEVELOPMENTS

The great advantage of this system is the possibility to alert people of critical episodes of air pollution in time to perform preventive actions to minimize or avoid the population to contact with the polluted air.

Before IQAr, preventive actions were impossible, because all data used to be compiled into a daily report up to

24 hours of delay. Thus, in case of criticals pollution events, the population would be inevitably exposed to a polluted air which might cause serious health damage.

Nowadays, IQAr continuously processes data from 13 air quality stations and this amount will possibly double in 2015. Prior to the implementation of IQAr, it was necessary to use various software for processing, validation, reporting and dissemination of data. There was no standardization about the format of the data received by the processing central, and, for some stations, the own manufacturer's software with its own standards was supposed to be utilized. Those characteristics hardly limited the agility of the process turned the dissemination of information in real-time into an impossible task.

IQAr enabled to automate the boring and costly process of collecting, processing and disseminating data - previously held by environmental engineers - so that these professionals could became able to focus on the work of supervision and evaluation of data, which the system cannot do on its own, because it requires the expertise of the professional in the area.

In Brazil, the IQAr was the first software to provide information on air quality in real-time using dynamic maps. In the state of Sao Paulo, by CETESB, data are disseminated in real-time in a geo-referenced way [16]. But the map is not dynamic, and thereby, all stations cannot be seen on a single map, forcing the user to click on each region and open a new static map for viewing it. In the State of Rio de Janeiro, the state environmental institute has also a solution using maps, but the information is not updated in real-time. By the time of this paper was being written, the most recent information was dated 11 days ago.

One of the future implementations that can be mentioned is the development of a mathematical modeling of pollutant dispersion module. With this feature it would be possible to simulate in real-time the dispersion of pollution (plume) emitted by industries and vehicles around the stations using the updated meteorological data from these stations.

In the short term, other implementations are planned like to provide data from IQAr to the AQICN system, so that information relating to Curitiba and other Brazilian cities monitored by IQAr can be used for comparison in relation to other urban centers of the world; and also to increase the level of detail and interactivity of the dynamic map in order to facilitate the use and access to information by society.

REFERENCES

[1] World Health Organization, “Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide”, 2005.

[2] Ö. Özden, T. Dögeroglu, and S. Kara, “Assessment of ambient air quality in Eskisehir, Turkey,” Environment International, vol. 34, pp. 678-687, 2008.

[3] C. A. Frondizi, “Air Quality Monitoring: Theory and Practice,” Rio de Janeiro: E-papers, 2008.

[4] K. K. Khedo, R. Perseedoss, and A. Mungur, “A wireless sensor network air pollution monitoring System,” International Journal of Wireless & Mobile Networks, vol.2, no.2, pp. 1-5, May 2010.

[5] J. Yadav, V. Kharat, and A. Deshpande, "Zadeh-Deshpande (ZD) fuzzy logic based formalism for linguistic description of air quality: A case study," International Conference on Informatics, Electronics & Vision (ICIEV), 2014, doi: 10.1109/ICIEV.2014.6850706

[6] W. Kaminski, J. Skrzypski, and E. Jach-Szakiel, "Application of Artificial Neural Networks (ANNs) to Predict Air Quality Classes in Big Cities," International Conference on Systems Engineering, ICSENG, Aug. 2008, doi: 10.1109/ICSEng.2008.14

[7] D. Mendez, A. J. Perez, M.A Labrador, and J.J. Marron, "P-Sense: A participatory sensing system for air pollution monitoring and control," IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 344-347, March 2011.

[8] R.H. Narashid and W. M. N. W. Mohd, "Air quality monitoring using remote sensing and GIS technologies," International Conference on Science and Social Research (CSSR), pp. 1186-1191, Dec. 2010, doi: 10.1109/CSSR.2010.5773713

[9] J. H. Liu, Y. F. Chen, T. S. Lin, D. W. Lai, T. H. Wen, C. H. Sun, and J. A. Jiang, “Developed urban air quality monitoring system based on wireless sensor networks,” In Fifth IEEE International Conference on Sensing Technology (ICST), pp. 549-554, Nov. 2011.

[10] E. Yaacoub, A. Kadri, M. Mushtaha, and A, Abu-Dayya, "Air quality monitoring and analysis in Qatar using a wireless sensor network deployment," 9th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 596-601, July 2013 doi:10.1109/IWCMC.2013.6583625

[11] U.S Environmental Protection Agency, “Guideline for reporting of daily air quality – pollutant standards index (PSI)”, December 1998.

[12] Environment Canada webpage. [Online]. Available from: http://www.ec.gc.ca/casaqhi/default.asp?Lang=En&n=065BE995-1 2014.10.08

[13] JavaScript Object Notation (JSON) [Online]. Available from: http://www.json.org/ 2015.01.07

[14] South Coast Air Quality Management District. AQMD. [Online]. Available from: http://www3.aqmd.gov/webappl/gisaqi2/VEMap3D.aspx 2014.10.07

[15] PostgreSQL. [Online]. Available from: http://www.postgresql.org/ 2015.01.07

[16] ESRI ArcGIS Server [Online]. Available from: http://www.esri.com/software/arcgis/arcgisserver 2015.01.07

[17] ArcGIS API for JavaScript [Online]. Available from: https://developers.arcgis.com/javascript/ 2015.01.07

[18] PHP Language [Online]. Available from: http://php.net/ 2015.01.07

[19] JFreeChart [Online]. Available from: http://www.jfree.org/jfreechart/ 2015.01.07

[20] JQuey [Online]. Available from: http://jquery.com/ 2015.01.07

[21] Envitech Ltd. [Online]. Available from: http://www.envitech.co.il/ 2014.10.09

[22] AQICN [Online]. Available from: website http://aqicn.org/map/ 2014.10.03

[23] CETESB System. [Online]. Available from: http://sistemasinter.cetesb.sp.gov.br/Ar/php/mapa_qualidade_rmsp.php 2014.10.09

[24] Gmaps [Online]. Available from: https://www.google.com.br/maps 2015.01.07

Mining Spatial and Temporal Movement Patterns of Passengers on Bus Networks

Chunjie Zhou*, Pengfei Dai†, Fusheng Wang‡, Renpu Li*
*School of Software, Ludong University, Shandong, China, Email: lucyzcj@gmail.com
†Institute of Network and Technology, ICT(Yantai), Shandong, China, Email: daipf@int-yt.com
‡Department of Biomedical Informatics, Stony Brook University, New York, USA, Email: fusheng.wang@stonybrook.edu

Abstract—**The analysis of human behavior is the basis of understanding many social phenomena. Accurate and reliable human movement pattern mining can lead to instructive insight to transport management, urban planning and location-based services (LBS). As one of the most widely used forms of transportation, buses can tell a lot of stories about people, including passenger demands, areas people are interested in crossing each day, and their travel patterns. Based on a large database from a real bus system, this paper aims to mine spatial and temporal movement patterns of passengers: evaluating traveling time of passengers, predicting number of passengers to estimate passenger demand and the crowdedness in the bus, and identifying attractive areas for passengers. There are major challenges for mining human movement patterns on bus networks: inhomogeneous, seasonal bursty periods and periodicities. In this paper, we take a Poisson process approach to model and evaluate traveling time of passengers, which can reflect the time features of individuals and activity cycles among areas. To overcome the challenges, we propose three prediction models and further take a data stream ensemble framework to predict the number of passengers. To obtain meaningful patterns of attractive areas, we provide a hierarchical clustering based approach to group spatiotemporally similar pick-up and drop-off points, as people's interests to these areas vary significantly on time, days and seasons. Our performance study based on a real dataset of five months' bus data demonstrates that our approach is quite effective: among 86,411 passenger demands on bus services, more than 78% of them are accurately forecasted.**

Keywords–movement pattern; traveling time; attractive areas; passenger demand; hierarchical clustering.

I. Introduction

Human activity patterns have received a certain amount of attention in recent studies. Analyzing human activity data to obtain structural information of humans has become an important means of studying social systems. Analyzing of individual banknotes and civil aviation traffic are notable studies in recent years. These studies have shown that individuals follow simple and reproducible patterns of mobility in several different manners [1]. Mining human movement patterns can help us to understand urban form and travel composition which can be used to support urban planning in terms of facility location and site selection. Meanwhile, it also helps us to explore passenger travel demand in different areas and in different time periods from transport management perspective, which is very useful for bus management. From LBS providers' perspective, the knowledge of passenger movement and behavior pattern can help to provide better tailored LBS, such as point of interests (POI) recommendation for a given time within appropriate scope.

As one of the most widely used mode of transport, bus can tell a lot of stories. It can tell not only road network traffic condition, but also areas people are interested in crossing in a day and their related travel patterns, such as traffic demand and their movement. Conventional bus information analysis tends to focus on road network travel time and average speed estimation. Indeed, as bus services are required by the same individual during his/her daily routine, it offers a proxy to capture individual human movement patterns [2]. This provides a unique opportunity for us to take advantage of bus information to discover spatial and temporal movement patterns of passengers, a major focus of this paper.

There are three major goals of this paper. 1) Evaluate traveling time length of passengers. In order to get the time features of individuals, we follow the Poisson process to evaluate the traveling time of passengers. 2) Predict the number of passengers to estimate passenger demand and congestion degree. When traveling with buses, travelers care about not only the waiting time, but also the crowdedness in the bus. Overcrowded bus may drive away the anxious travelers and make them reluctant to take buses. We propose a novel methodology to produce online predictions on the passenger demand using time series forecasting techniques. 3) Identify attractive areas. Attractive areas are places that people often visit, for instance, hot shopping and leisure places or living and working areas based on their level of attractiveness (LoA). We take bus pick-up and drop-off points as the focus in this study, because they can convey rich information on identifying attractive areas and associated movement patterns. An area's LoA can be determined by travelers' visiting frequency, which can be measured by the number and density of bus pick-up and drop-off points.

Accurate, real-time and reliable human movement pattern mining is the basis of understanding many social phenomena. However, due to a number of stochastic variables, we need to face the following three major challenges: 1) *inhomogeneous*. A periodicity in time on a daily basis that reflects the patterns of the underlying human activity, making the data appear non-homogeneous. 2) *seasonal bursty periods*. The movement patterns of passengers can be often messed by seasonal bursty periods of expected events such as highly crowded holiday events, weather changes, and so on. 3) *other periodicities*. The passengers' demands and attractive areas vary significantly at different time of a day, different day of a week, or even different seasons.

Aiming to address these challenges, in this paper, we present methods to mine spatial and temporal movement patterns of passengers on a bus network. To predict passenger demands, we develop a unique predictive model by adapting time series forecasting techniques to our problem. To discover attractive sites, we develop a hierarchical clustering based approach to group similar pick-up and drop-off points since people's interests to these areas varies through time of the day, day of the week, even season of the year.

In our work, we have conducted a real study using a dataset obtained by a large-sized bus network containing a total of 416 bus stops and 1,326 vehicles running in the City of Yantai, China. Our test-bed is a computational stream simulation running offline. The data from the first 16 weeks were used as training set and the data from the last 6 weeks were used as input for our stream-type test-bed, i.e., simulating the movement patterns that would arrive continuously in a stream. Our experiments demonstrated promising results of our approach: our model can accurately predict more than 78% of actual passenger demands for bus services.

Contributions. The major contributions of the paper are summarized below.

- We take a Poisson process approach to model and evaluate traveling time of passengers, which can reflect the time features of individuals and activity cycles among this area.
- We propose three distinct prediction models and a well-known date stream ensemble framework to predict the number of passengers to estimate passenger demand and congestion degree. These three models can gradually solve the challenges of inhomogeneous, seasonal bursty periods and periodicities.
- We provide a hierarchical clustering based approach to identify attractive areas by grouping similar pick-up and drop-off points, since people's interests to these areas varies significantly at different time of a day, different day of a week, or even different seasons.
- Our comprehensive comparative performance study based on real datasets demonstrates the effectiveness of our methods.

Paper Organization. The rest of the paper is organized as follows. Section II summarizes the related work. Section III introduces the time features of passengers. Section IV proposes three distinct prediction models and a well-known data stream ensemble framework to predict the number of passengers. In Section V we introduce our study method to identify attractive areas. Section VI describes the experiments based on a real-world scenario, including the evaluation metrics of our model, the experimental setup, and the experiment results. Section VII concludes the paper and describes the future work.

II. Related Work

Some studies [3] utilized data mining approaches, such as clustering to extract meaningful information by analyzing trajectory stops, moves and their sequences. Work in [4] attempted to address the problem by introducing semantic modeling process into trajectory points so that meaningful patterns can be extracted from trajectories, as background geographic information is of fundamental importance for both traffic-oriented and user scenario-oriented analysis. Otherwise, patterns may be incomplete due to the missing of POI [3].

Semantic model approach [5] combines background geographic information together with trajectory location (x, y), and the concepts of stops and moves [6] are often used to facilitate discovering and modeling trajectory patterns. However, semantic modeling approach is only applicable when trajectory points can be matched with background geographic places precisely. Since trajectories are often matched onto road centerlines, while POIs are distributed along road links and stored as point object in map database, this approach may not be applicable under many circumstances unless POIs are represented by polygons and trajectory points are dense enough to surround these POIs.

To generate meaningful patterns, various types of clustering techniques have been popularly used. Trajectory point density, frequency, and stay time are the most frequently used factors to assist information extraction. For example, Alvares et al. [5] extracted moving patterns which were assumed to follow Markov chain between general type of stops, such as hotels, airports and tourist places. It also assumed each stop was located at a POI, and used travel frequency to judge their "importance". Palma et al. [3] identified stops by using a density-based clustering algorithm and introducing a "minimal stop durations" which takes into account of the average periodicity of the trajectory time points. Li et al. [4] further considered users' travel experiences, and discovered the association among these points, for instance, classical travel sequence. Verhein and Chawle [7] defined spatio-temporal association rules and related concepts and found patterns using pruning properties based on synthetic dataset.

Different from the above studies which are mostly based on limited amount of personal trajectories, our study uses large amount of bus information data to explore time-dependent attractive areas and movement patterns. Such patterns can be represented by high traffic demand areas and passenger movement among them. This is more complex in terms of wider geographic coverage and diverse individual trip purposes. Since bus pick-up and drop-off points represent traffic demand, clustering these points is a feasible approach to discover areas with high travel demand. During this process, travel interactions among the clusters and other information are obtained. The detailed approach will be illustrated in the following sections.

III. Traveling Time of Passengers

As one of the main parts of mining spatial and temporal movement patterns of passengers, we first put all the buses as a whole to study the traveling time of passengers. Based on this result, we may roughly get the time features of individuals and activity cycles among this area. We follow the Poisson process to predict the traveling time and assume that at a given interval time t, the probability of n-incident occurred is:

$$P(n,q) = e^{-qt}\frac{(qt)^n}{n!} \tag{1}$$

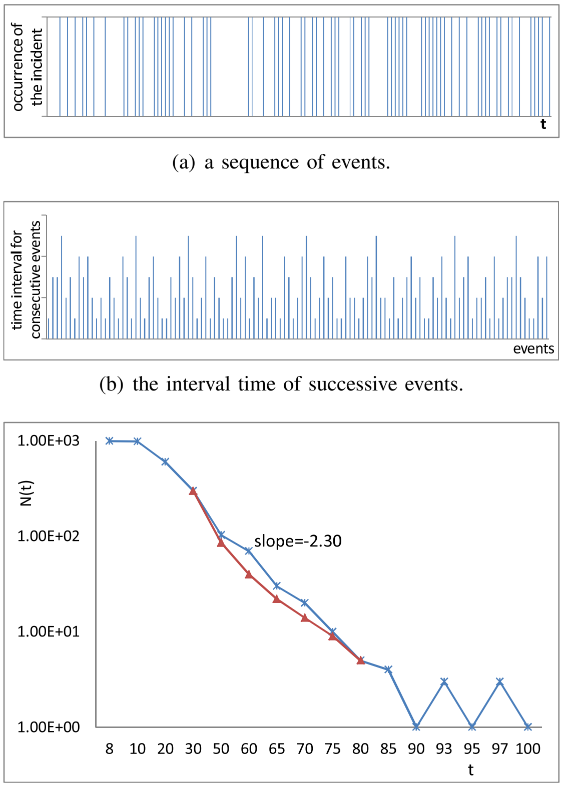
Here, q represents the probability of an event occurrence. Based on (1), we can derive that the interval distribution of two consecutive events is:

$$P(\tau) = qe^{-q\tau} \tag{2}$$

Here, τ means the traveling time of two consecutive events. To test this hypothesis, we analyze the data of traveling time and find that the distribution of traveling time over all buses is well approximated by a Power-law:

$$P(\tau)\infty\tau^{-\alpha} \tag{3}$$

Here, the exponent is between 2.20 and 2.90. We also find that the traveling time of passengers are mainly in the range of 15-30 minutes, and only a small portion of intervals are longer



(a) a sequence of events.

(b) the interval time of successive events.

(c) the distribution of data points.

Figure 1. The distribution of the traveling time.

than an hour. The results of traveling time analysis are shown in Figure 1.

Figure 1 shows the traveling time distribution of passengers. Figure 1(a) shows a sequence of events. Each vertical line means an occurrence of the incident at that time. Figure 1 (b) shows the interval time of 200 successive events, and the length of the vertical line corresponds to the time interval for two consecutive events. Figure 1(c) shows the distribution of part of the data points in (a) and (b) diagrams in double logarithmic coordinates.

IV. The Passenger Demand Prediction

Our goal is to predict the number of passenger demands of a certain bus b at the bus stop s at instant t. To achieve this, we have studied three distinct prediction models, and a data stream ensemble framework.

A. Time Varying Poisson Model

The passenger demand on bus services normally follows a periodic pattern on a daily basis.

Here, we assume the probability $P(n)$ of n buses appearing in a determined time period follows a Poisson distribution. We can define it using the following equation:

$$P(n; \lambda) = \frac{e^{-\lambda}\lambda^n}{n!}, \tag{4}$$

Here, λ represents the rate (averaged number of passenger demands on bus services) in a fixed time period. However, in this specific problem, the rate λ is not constant but time-variant. Thus, we adapt it as a function of time, i.e., $\lambda(t)$, and transform the Poisson distribution into a nonhomogeneous one. Let $\lambda(t)$ be defined as follows:

$$\lambda(t) = \lambda_0 \delta_{d(t)} \eta_{d(t),h(t)}, \tag{5}$$

Here, $d(t)$ represents the weekday $\{1 = Sunday, 2 = Monday, ...\}$; $h(t)$ is the period in which time t falls in.

The model requires the validity of both equations

$$\sum_{i=1}^{7} \delta_i = 7, \tag{6}$$

$$\sum_{t=1}^{T} \delta_{t,i} = T \quad \forall t, \tag{7}$$

where T is the number of time spans in a day. To ease the interpretation of these equations, we can define the remaining symbols as follows:

- λ_0 is the average rate (i.e., expected rate) of the Poisson process over a full week;
- δ_i is the relative change for the day i (Saturdays have lower day rates than Tuesdays);
- $\eta_{j,i}$ is the relative change for the period i on the day j (the peak hours);
- $\lambda(t)$ is a discrete function representing the expected distribution of passenger demands on bus services over time for a bus stop of interest s.

B. Weighted Time Varying Poisson Model

The model above can predict the time-dependent average number of passenger demands on bus services. However, it is not guaranteed that every bus stop has highly regular passenger demands: indeed, the demands in many stops can be often messed by seasonal bursty periods of expected events such as highly crowded holiday events, weather changes, and so on.

To tackle this special seasonal issue, we propose a weighted average model based on the above presented approach. Our goal is to increase the relevance of the demand pattern observed in the last week comparing to the patterns observed several weeks ago.

Here, the weight set w is calculated using a well known time series approach – the Exponential Smoothing approach [8]. We define w as follows:

$$w = \alpha * \{1, (1-\alpha), (1-\alpha)^2, ..., (1-\alpha)^{\lambda-1}\}, \tag{8}$$

Here, λ is the number of historical periods considered in the initial average, α is the smoothing factor (i.e., a user defined parameter) and $0 < \alpha < 1$.

C. Autoregressive Integrated Moving Average Model

The last two models assume the existence of a regular (seasonal or not) periodicity in passenger demands on bus services. However, there are other periodicities, for example, the number of passenger demands in one bus stop of a certain bus in regular workdays during a certain period is highly similar. Moreover, the number of passenger demands on bus services in the morning and in the evening of the same day is also very similar.

ISBN: 978-1-61208-383-4

We explore the Autoregressive Integrated Moving Average Model (ARIMA) [10]. In this model, the future value of a variable is assumed to be a linear function of several past observations and random errors. We can formulate the underlying process that generates the time series (passenger demands on bus services over time for a given bus stop s) as:

$$\begin{aligned} R_{s,t} &= \theta_0 + \phi_1 X_{s,t-1} + \phi_2 X_{s,t-2} + ... + \phi_p X_{s,t-p} \\ &+ \varepsilon_{s,t} - \theta_1 X_{s,t-1} - \theta_2 X_{s,t-2} - ... - \theta_q X_{s,t-q} \end{aligned} \quad (9)$$

Here, $R_{s,t}$ and $\varepsilon_{s,t}$ are the predicted value and the random error at time period t respectively; $\phi_l (l = 1, 2, ..., p)$ and $\theta_m (m = 0, 1, 2, ..., q)$ are the model weights; p and q are positive integers often referred as the orders of the model. Orders and weights can be inferred from the historical time series using both the autocorrelation and partial autocorrelation functions.

D. Sliding Window Ensemble Framework

We have proposed three distinct predictive models to learn from long, medium and short-term historical data. Then, how can we combine them all to further improve our prediction?

Let $M = \{M_1, M_2, ..., M_z\}$ be a set of z models of interest to model a given time series and $M_t = \{M_{1t}, M_{2t}, ..., M_{zt}\}$ be the set of forecasted values to the next period on the interval t by those models. The ensemble forecast E_t is obtained as

$$E_t = \sum_{i=1}^{z} \frac{M_{it}}{\beta}, \beta = \sum_{i=1}^{z} \rho_{iH} \quad (10)$$

where ρ_{iH} is the forecasting accuracy obtained for the model M_i in the periods contained in the time window $[t - H, t]$. H is a user-defined parameter to define the window size. As the information is arriving in a continuous manner for the next periods $\{t, t+1, t+2, ...\}$, the window will also slide to determine how are the models performing in the last H periods. To estimate such accuracy, we take a time series forecasting error metric: the Symmetric Mean Percentage Error (sMAPE) [11].

V. Discovery of Attractive Areas of Passengers

Since each pick-up and drop-off stop can only represent the approximate location where travel demand generates from, it is more proper to represent LoA using a polygon instead of a point. Clustering is a feasible method to eliminate noises in discovering the patterns while defining the borders of attractive areas. After pick-up and drop-off points are clustered, passenger movement pattern can be identified more clearly with more information, such as flow interaction and average travel distance.

Since the number of clusters cannot be known beforehand, we take a hierarchical or agglomerative clustering algorithm instead of partition clustering. An example of partition clustering is k-mean, which requires a pre-knowledge of cluster number and the shapes of all clusters have to be convex. We follow the notion that the distance $dist(i, j)$ between two points (i, j) measures their dissimilarity and determines the possibility as a cluster. Here, we adopt Euclidean distance. Some studies consider network constraint as an improvement in trajectory clustering, however, in this study, as the interest is in "area", Euclidean distance is more proper.

We take the single-linkage or nearest neighbor clustering criterion, which is one of the most widely used clustering criterias:

$$Dist(c_n, c_k) = Min(Dist(c_i, c_k) + Dist(c_j, c_k)) \quad (11)$$

where two clusters C_i and C_j are merged to generate a new higher level cluster C_n, and C_k is the remaining cluster.

Algorithm 1 Clustering Algorithm

```
Require:
  matrix $D$ contains all distances $d(i, j)$.
Ensure:
  $L(k)$ is the level of the $k^{th}$ clustering;
  the proximity between clusters $(n)$ and $(k)$ is denoted as $d[(n), (k)]$.

  Level $L(0) = 0$;
  sequence number $m = 0$;
  while objects in more than one clusters do
    Find the least dissimilar pair of clusters in the current clustering;
    $d[(n), (k)] = mind[(i), (j)]$;
    Cluster($m$)=cluster($n$) + cluster($k$);
    Merge clusters $(n)$ and $(k)$ into a single cluster to form the next clustering $m$;
    $L(m) = d[(n), (k)]$;
    $m = m + 1$;
    Update matrix $D$;
  end while
```

Hence, the distance between two clusters is computed as the distance between the two closest elements in the two clusters. As a result, clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other. This is called as "chaining phenomenon", and is usually treated as a drawback of single-linkage method. However, the bus information is constrained by linear road links which makes the most distinctive feature comparing with other spatially distributed data. Conventional clustering algorithms, even the emerging spatio-temporal clustering method [9] do not take linear distribution into consideration. Based on the notion that vehicles driving on the same road links usually share similar destination, trajectories along the same road link thus are more similar compared with those distributed on different road links even if with smaller spatial distance. Therefore, more weight is given to the trajectory points on the same road link, which exactly conforms to the chaining phenomenon.

Moreover, single-linkage clustering is deterministic, in the sense that the resulting clusters do not depend on the order in which elements having equal distances are chosen. This is not necessarily true of other linkage schemes. The clustering algorithm is described in Algorithm 1.

VI. Experimental Results

In this section, we first describe the experimental setup developed to test our model on the available data, then present and discuss the results.

A. Experimental Analysis of Passenger Demands

Our model produces an online forecast for the passenger demands in all bus stops at each P-minutes period. Such test is through an offline continuous simulation. The scripts used are developed using the R statistical software. The predefined functions used and the values set for the model's parameters are detailed along this section.

TABLE I. ACCURACY OF MODELS USING ALPHA(α)=0.4.

Model	Periods			
	5am to 9am	9am to 1pm	1pm to 5pm	5pm to 9pm
Poisson Mean	75.83%	71.69%	74.13%	73.58%
W. Poisson Mean	76.98%	74.27%	75.76%	75.18%
Arima	78.35%	73.79%	76.28%	75.99%
Ensemble	79.26%	76.59%	78.39%	77.96%

TABLE II. ACCURACY OF MODELS USING ALPHA(α)=0.5.

Model	Periods			
	5am to 9am	9am to 1pm	1pm to 5pm	5pm to 9pm
Poisson Mean	75.83%	71.69%	74.13%	73.58%
W. Poisson Mean	76.63%	73.58%	74.86%	74.39%
Arima	78.35%	73.79%	76.28%	75.99%
Ensemble	79.37%	76.62%	78.33%	78.06%

Here, the aggregation period is set to 30 minutes (i.e., a new forecast is produced each 30 minutes; P=30) and a radius of 100 meters is used ($W = 100$). These parameters are set according to the average waiting time in bus stop ($<$ 21 minutes).

The previously described dataset is divided into a training portion of 16 weeks, and a test portion of 6 weeks. Both training and test set are composed by one time series per bus stop and each value tested on the period t is merged to the training set to generate the forecast on the period $t+1$, i.e. the real number of passenger demands count for each bus at a given bus stop along 30 minutes are considered for the next period forecast, and so on.

The time-varying Poisson averaged models (both weighted and non-weighted) are updated every 24 hours. A sliding window of 4 hours (H=4) is considered in the ensemble. The accuracy of each model is measured using the metric, which is also used to weight each model in the ensemble - *sMAPE*. Distinct results for two distinct values (0.4 and 0.5) of the parameter *alpha* (α in the weighted average) are presented below.

A total of 86,411 bus services are tested. The accuracy measured for each model is presented in Table I and Table II. The results are firstly presented per shift and then globally. The values presented below are calculated through an average weight of the accuracy obtained in each one of the time series (i.e., the accuracy of the forecast on the passenger demand for buses on each one of the 416 bus stops). Each accuracy is weighted according to the number of services demanded on the corresponding bus stop along all the test periods.

Each model presents accuracy above the 76% in both tables. The W. Poisson Mean and the Ensemble are the only ones affected by the changes on the *alpha* (α) parameter. The sliding window ensemble is always the best model in every shift and period considered, with an accuracy superior to 78%: 67,400 of the 86,411 total services are correctly forecasted in both time and space using an aggregation of 30-minutes periods.

TABLE III. CLUSTERING RESULTS ON MONDAY.

Time Span	Total points	No. of clusters (n>60)
5am-9am	10,682	12
9am-1pm	7,205	32
1pm-5pm	9,420	28
5pm-9pm	8,793	21

TABLE IV. CLUSTERING RESULTS ON SATURDAY.

Time Span	Total points	No. of clusters (n>60)
5am-9am	5,481	36
9am-1pm	7,915	28
1pm-5pm	9,506	15
5pm-9pm	10,472	21

B. Experimental Analysis of Attractive Areas

To discover attractive areas, or so-called hot regions, we choose those areas that have over 60 pick-up and drop-off points. This is because each time span is 4 hours, i.e., 240 minutes. The 60 points can ensure there emerges at least one pick-up or drop-off point in the cluster less than every 4-minute in average, which is dense enough. As described above these vehicles usually run in one out of four 4-hour shifts: 5am-9am, 9am-1pm, 1pm-5pm and 5pm to 9pm. Each time span corresponds to a cluster result, as summarized in Table III and Table IV. Table III shows the cluster results on Monday, while Table IV shows the cluster results on Saturday. From the total number of pick-up and drop-off points and the number of clusters, we can see that on a typical workday, people tend to be more active during the time span of 5am-9am and 1pm-5pm, while in a typical weekend, people tend to be more active in the afternoon and at night.

Figures 2 and 3 give an overview of the distribution of studied bus pick-up and drop-off points on workdays and on weekends separately. From the distribution of the clusters in two figures, it can be further observed that the distribution of the attractive areas follow certain pattern while vary crossing workdays and weekends. For example, there are some areas which remain attractive, though the sizes of them vary with time. There are also areas whose LoA are time-dependent. Moreover, we can see that people's activities are more concentrated on work places and living places on workdays, while places of amusement and schools are more attractive on weekends.

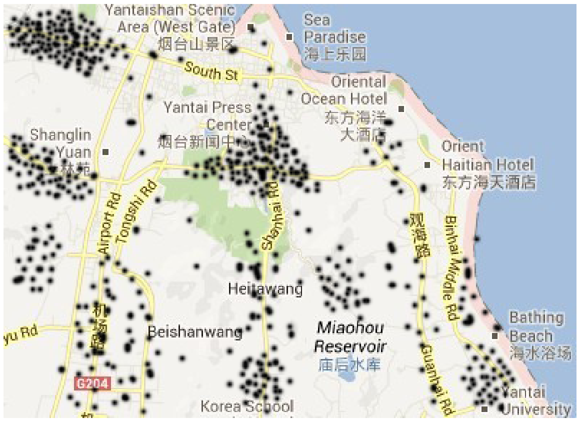


Figure 2. The cluster distribution on workday.

C. The Analysis of Movement Patterns of Passengers

The billing system of bus services records each passenger's behavior when a passenger gets on a bus and gets off a bus.

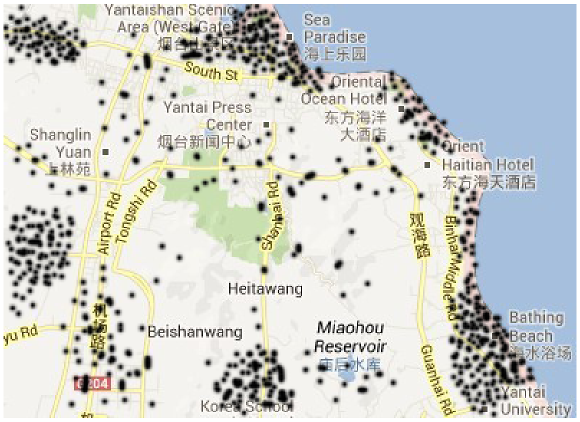


Figure 3. The cluster distribution on weekend.

Therefore, there is a certain correlation between the data of buses and the behavior of people in bus services. In order to study the correlation, we explore the data of buses to find the behaviors of people. The foregoing analysis shows that people who take a bus service may have the following characteristics:

- The service time is mainly about 15-30 minutes and only a minority is longer than an hour. The distribution of service time interval is approximated by a Power-law.
- In 77% cases, the number of passengers is over the available capacity of bus services. That means the buses are too crowded in most cases.
- Passengers' active time is often much concentrated. People tend to be more active during the time span of going to work and getting off work on workdays, while on weekends, people tend to be more active in the afternoon and at night.
- Attractive areas are time-dependent. People's activities are more concentrated on work places and living places on workdays, while places of amusement and schools are more attractive on weekends.

VII. Conclusion

In this paper, we present our work on mining spatial and temporal movement patterns of passengers on bus networks from three aspects: evaluating traveling time of passengers, predicting number of passengers to estimate passenger demand and congestion degree, and identifying attractive areas. The study is performed by transforming both GPS and event signals emitted by 1,326 buses in the City of Yantai, China into time series of interest. As a result, our method is able to identify attractive areas and predict passenger demand on buses at each one of the 416 bus stops at every 30-minutes.

Our method demonstrates a satisfactory performance, and predicts accurately more than 78% of the 86,411 demanded services, anticipating in real time the spatial distribution of the passenger demand. The approach is novel and can provide instructive insight to transport management, urban planning and location-based services. In particular, our work has major practical impact to help the management of bus companies to provide optimal services based on the knowledge of passenger movement patterns.

Acknowledgment

This research was partially supported by the Project of Shandong Province Higher Educational Science and Technology Program (No.J12LN05); the grants from the National Natural Science Foundation of China (No.61202111, 61273152, 61303017); the Project Development Plan of Science and Technology of Yantai City (No.2013ZH092); the Doctoral Foundation of Ludong University (No.LY2012023); the natural science foundation of Shandong province and special projects (Grant No.ZR2013FL009); the US National Science Foundation ACI 1443054 and IIS 1350885; Scientific Research Foundation for Returned Scholars of Ministry of Education of China (43th).

References

[1] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi, "Understanding Individual Human Mobility Patterns". Nature 453, 2008, pp.779-782.

[2] M. Jordan, J. Kleinberg, and B. Scholkopf, "Pattern Recognition and Machine Learning[C]". Christopher M. Bishop. 2006.

[3] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A Clusteringbased Approach for Discovering Interesting Places in Trajectories". in SAC'08 Fortaleza, Ceara, Brazil, 2008, pp. 863-863.

[4] X. Li, C. Claramunt, C. Ray, and H. Lin, "A Semantic-based Approach to the Representation of Network-constrained Trajectory Data". In 12th International Symposium on Spatial Data Handling: Springer, 2006, pp. 451-464.

[5] L. O. Alvares, V. Bogorny, J. F. Macedo, B. Moelans, and S. Spaccapietra, "Dynamic Modeling of Trajectory Patterns using Data Mining and Reverse Engineering". In 26th International Conference on Conceptual Modeling - ER2007- Tutorials, Posters, Panels and Industrial Contributions. vol. 83, 2007, pp. 149-154.

[6] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. Macedo, F. Porto, and C. Vangenot, "A Conceptual View on Trajectories". Data and Knowledge Engineering vol. 65, 2007, pp. 126-146.

[7] F. Verhein and S. Chawla, "Mining Spatio-temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases". In 11th International Conference on Database Systems for Advanced Applications. vol. LNCS 3882, M. L. Lee, K. L. Tan, and V. Wuwongse, Eds.: Sringer-Verlag, 2006, pp. 187-201.

[8] C. Holt, "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages". International Journal of Forecasting, 20(1), 2004, pp.5-10.

[9] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data". Data and Knowledge Enginnering. Vol 60, 2007, pp. 208-221.

[10] L. H. B.Williams, "Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results". Journal of Transportation Engineering, 129(6), 2003, pp.664-672.

[11] S. Makridakis, and M. Hibon, "The M3-Competition: Results, Conclusions and Implications". International Journal of Forecasting, 16(4), 2000, 451-476.

Geoprocessing Applied to Open Government Data

Daniel Farias Batista Leite, Julio Henrique Rocha,
Claudio de Souza Baptista, Ana Gabrielle Ramos
Falcão
Federal University of Campina Grande, Campina Grande,
Paraíba, Brazil
{danielfarias, juliorocha}@copin.ufcg.edu.br,
baptista@dsc.ufcg.edu.br anagabriellee@gmail.com

Hugo Feitosa de Figuerêdo
Federal Institute of Education, Science and Technology of
Paraiba, Monteiro, Brazil
hugo.figueiredo@ifpb.edu.br

***Abstract*—The increasing social supervision in various sectors of society, particularly in the government administration, has demanded for information access policies in several countries. In Brazil, for instance, there is a law that regulates access to the Open Government Data (OGD) for reading, monitoring and reusing data in new projects and applications. In this scenario, several applications that extract information from open data are being developed throughout the world. This paper presents the *BrasilMaps* system, an open source application that integrates, through GIS technology, twelve open data government services offered to the Brazilian citizens. Data as Basic Health Units (BHU), Notary Offices, National Employment System (SINE), Service Unit of the Inland Revenue, Universities and Federal Institutes, Agencies of the Ministry of Labor and Employment and the Federal Police Offices are made available in the proposed system. The proposed system provides information that helps people in the use of public services. Through theaccess to this data, citizens can benefit from the nearest services offered, identify the quality of some services and filter the desired services.**

Keywords- Open Government Data; Geoprocessing; GIS.

I. Introduction

Over the last years, geoprocessing technologies have played an important role in the provision of solutions for the public sector. GIS presents a great potential for information dissemination, reflecting in several sectors of the society, especially in the public sector, in which we can note an intense social control. The use of GIS enhanced with social media enables a higher supervision and participation of the population in government policies.

As a consequence of such social control in the public sector, many countries have invested in information access policies, such as the United States and New Zealand. In Brazil, in particular, two of these laws stand out: the Complementary Law 131/2009, which states the transparency of the financial execution of the Union/State/City; and the Public Information Access Law 12.587/2011.

The Public Information Access Law (IALaw) establishes the access to Open Government Data (OGD), which are data produced by the government and made available for the citizens in order to enable not only its monitoring and accessing, but also its reuse in new projects, websites and applications. Furthermore, the government data are only considered open when published in conformity to eight principles: complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary and license-free [1]. In addition there are three important laws [2]:

- whether the data cannot be found or indexed on the web, it does not exist;
- if it is not open and available in machine readable format, it cannot be reused; and
- if any legal framework does not allow it to be repurposed, it is not useful.

The adoption of Open Government Data policies is beneficial in many ways. At first, with the transparency of the government actions, one can create a collaborative social control environment for both the population and the government. Matheus et al. [3] also highlight another benefit of the adoption of Open Data, which is the possibility of generating new information and applications from the interaction between the government and the society through the exploration of open data derived from several data sources.

In this paper, we propose anGIS application named ***BrasilMaps***[4] , that enables the integration of heterogeneous data sources from twelve data bases of public equipments (extracted from the Open Data National Portal - "Portal Nacional de Dados Abertos"), using geoprocessing techniques and allowing visualization and searching on such data.

The database containsdata on Basic Health Units, notary offices, Employment National System, Internal Revenue Service, *Fundacentro* (responsible for the registration of work accidents and diseases), Universities and Federal Institutes, Ministry of Labor and Employment and Federal Police offices. It also contains data about services provided for assisting individuals in threat or right violation situations (physical, psychological or sexual violence, amongst others), such as the Social Assistance Specialized Reference Center the Social Assistance Reference Center, Therapeutic Communities (specialized in sheltering people with disorders derived from psychoactive substances) and Social Assistance Private Networks.

In order to provide a good visualization on the map of the aforementioned sources, it is necessary to use a methodology based on the Extract, Transform and Load (ETL) process.

Therefore, the data sources will be extracted and examined so that only the ones that qualify are loaded and, if necessary, normalized. In many data sources, we can find incomplete and uncertain information due to several factors that may occur during the data collection.

When using many data sources, some challenges may rise on the data treatment process, notably: normalization of the geographic information, information filtering, elimination of imprecise information, correction of uncertainties and errors, incomplete information and outdated data. Such challenges were overcome using a methodology for the data treatment obtained from the aforementioned sources, thus facilitating the exhibition of the information in the map.

The rest of this paper is organized as follows. In Section II, we discuss related work and applications that make use of Open Government Data. Next, we address the methodology used for obtaining and manipulating the open data in Section III. In Section IV, we focus on the ***BrasilMaps*** application in details, including its architecture and available functionalities. In Section V, we describe the usability assessment performed to evaluate ***BrasilMaps***, and finally, in section VI, we present our conclusions and the future work.

II. Related Work

It is understandable that, due to the short lifetime of the IALaw, the Open Government Data policy in Brazil is still under the process of development and improvement. In this sense, few applications like ***BrasilMaps*** were been developed. Mootiro Maps [5]works with five public services datasets and shows all the information in only one layer at map, so it is impossible to filter information. Moreover, a single icon is used to represent all the public services. PAC II [6] displays the location of public constructions but does not distinguish the icon according to the category of construction, only a single icon is used.

Pedroso et al. [7] highlight several problems that some datasets made available by the government agencies, such as their availability in only the *.pdf* format, which makes interoperability difficult to achieve. In this perspective, Lourenço and Serra [8] defined a maturity model composed of 5 levels, in which the level 0 indicates the available databases that do not present any structure, whereas the highest maturity level (level 4) indicates that the available information is structured.

Breitman et al. [9] performed a quantitative and qualitative analysis of the datasets available in the Brazilian Open Data National Portal, indicating the categories the datasets are organized and the total volume of each category, the origin of each dataset (spreadsheets, web services, among others) and the format they are made available (XML, Json, among others). On the other hand, Xu and Zheng [10] performed an analysis on the open data available on the world under the user perspective, indicating which categories the population would like to have access to and how such data should be made available.

Matheus et al. [3] and Correa et al. [11] performed a comparative study between several Brazilian transparency portals, regarding the federal, state and city levels, coming to the conclusion that, none of the portals are in conformity to the eight aforementioned principles for data to be considered open. Furthermore, Erickson et al. [12]argue that applications developed using open data should be regulated according to some rules, such as the way the data must be explored.

Even in such unfavorable scenario, however, many applications using open data information has been developed in Brazil. Nonetheless, these related applications use only a single database and, most of them, do not explore the spatial dimension of data. They only display the information in the form of charts and tables. Artigas [13] and Artigas and Chun [14] highlight the importance of exploring other means for displaying data, especially geolocation. The exploration of the spatial dimension may enable the detection of tendency and patterns between datasets. Graves and Hendler [15] state that displaying the data in many forms is essential for the entertainment of the citizens. In this context, this paper presents the ***BrasilMaps*** application that apart from integrating data from twelve open databases, enables the spatial visualization and interpretation of the data.

III. Data Transformation Process

The ***BrasilMaps*** application performs the spatial integration of twelve databases. However, in order to accomplish the integration of these heterogeneous databases a data transformation process was necessary, once some databases did not have geolocation information of the underlying equipments. Figure 1 illustrates the data transformation process implemented.

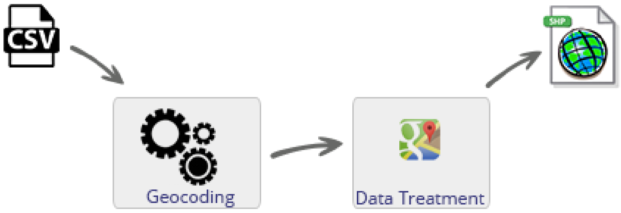


Figure 1. Extract, Transform and Load (ETL) Process.

All the files were collected using the comma-separated values format (.csv), thus being in conformity to the technical rules for the publication of Open Data, other than preserving the original structure of the information. Then, these files were submitted to the geocoding module. The geocoding module is responsible for performing an analysis on the geographical information of the databases and converting them into geographic coordinates, in case they are in the form of addresses. This conversion is called Geocoding, which is the process of identifying the geographic coordinates (e.g., latitude and longitude) from other geographic information, such as an address or a postal code. In ***BrasilMaps***, the geocoding process is performed by a web service available from *Google Maps*.

It is known that it is very unlikely to achieve a full conversion of the address information into geographic coordinates. This is due to many reasons, such as the absence of complementary information (postal code, neighborhood), the misspelling of street names, and others; and all this data that were not successfully geocoded were discarded. It is better not to display them than displaying them erroneously. At the end of the refining process, about 92% of the submitted data were successfully converted. The failures encountered happened due to: incomplete address information of the public service; typos or misspellings from the team responsible for gathering the open government data; and outdated addresses. At the geocoding stage, we must mention the high success rates obtained for the data related to SINE and *Fundacentro*. Figure 2 presents the success and fail rates for each database considered, making it possible to note the low success rates of the other databases, such as the Federal Revenue information: only 77%.

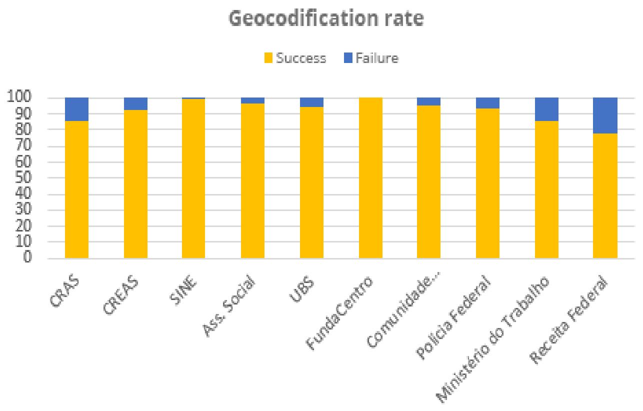


Figure 2. Geocoding success rates.

After the refining process, in which all databases presented associated geographic coordinates (latitude and longitude), it is necessary to convert them into spatial objects compatible with the spatial Database Management System chosen: PostgreSQL/PostGIS. The spatial data conversion module is responsible for this task, performing the instantiation of a new spatial object through PostGIS's *ST_Geom_FromText* function.

Finally, after the data transformation process we used Java Database Connectivity (JDBC) for sending SQL instructions to the database, loading the spatial database with the data gathered, considering each database as an entity responsible for storing equipment specific information. Next, we used *OpenLayers* to import and publish the entities with the support of the map server *GeoServer*.

IV. BRASILMAPS

This section introduces ***BrasilMaps***, an application that allows the visualization, in the spatial dimension, of the government services offered to the population. This system was developed using the programming language *JavaScript* and *HTML5*, thus allowing its access via web navigators, even from mobile devices. ***BrasilMaps*** was awarded with the first prize in the 2ndBrazilian National Open Data Competition.

The main goal of ***BrasilMaps*** is to provide a Geographic Information System (GIS) that enables the loading, processing and displaying of the Open Data for the population (made available from the Federal Government), offering a simple and practical public service search mechanism in a region of interest. The system's ability to categorize each public service results in a potentially higher and more effective participation of the society in consuming the open information that until then was of hard access to.

BrasilMaps makes use of the user's current location to display the public interest services locations closer to where he is or to any other given location. User location is detected from the Internet once it has been authorized by the user, as Figure 3 shows. When the user accesses the system from a Global Positioning System (GPS) enabled mobile device, a higher precision on the detection of the current location is achieved. This way, when the access to ***BrasilMaps*** is done in a big city (São Paulo/SP, for instance), the accurate location detection assists the user when performing searches and also results in faster responses from the system, once only the public service within the user's region of interest will be shown.

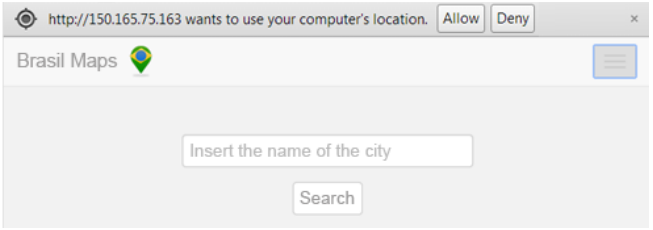

http://150.165.75.163 wants to use your computer's location.

Figure 3. ***BrasilMaps*** initial page.

Figure 4 illustrates the moment of the user's geographic location detection when accessing ***BrasilMaps*** from a mobile device.

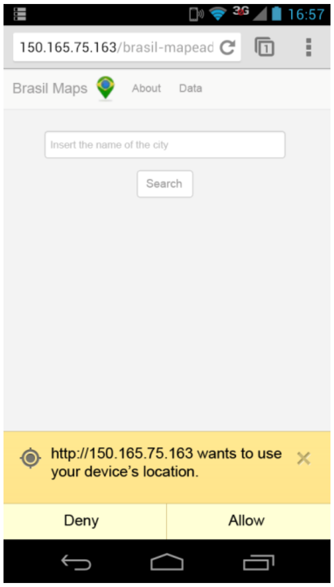

http://150.165.75.163 wants to use your device's location.

Figure 4. Accessing ***BrasilMaps*** from a smartphone.

In both situations, in case the user chooses not to share his location, a search field is provided for him to insert any address or city desired. This way, the application is launched with the map centered on the location submitted.

The following subsections describe ***BrasilMaps*** architecture and its functionalities details.

A. *Architecture*

BrasilMaps was designed according to the 3-tier model, aiming to facilitate portability, remote data updating and system modularization. The three layers that composed the system are: the Application Layer, the Business Layer and the Data Layer. Figure 5 illustrates the system's 3-tier architecture.

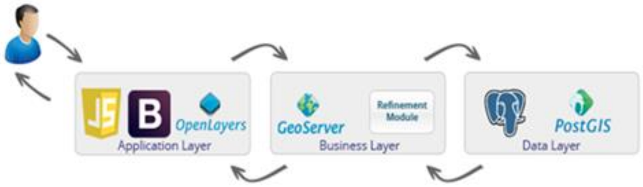


Figure 5. ***BrasilMaps*** architecture.

The Application Layer, responsible for presenting the information for the final user, provides mechanisms for loading, managing and viewing the data in a simple and intuitive mode. This layer has all the functionalities necessary for the users to benefit from the information of their interest. The Application Layer is responsible for communicating with the Business Layer performing requests, handling responses and formatting data properly to correctly display the information to end users.

The Business Layer is the intermediate element between the Application Layer and the Data Layer. This layer is responsible for managing the flow of information consumed and forwarded by the application in terms of requests and responses.

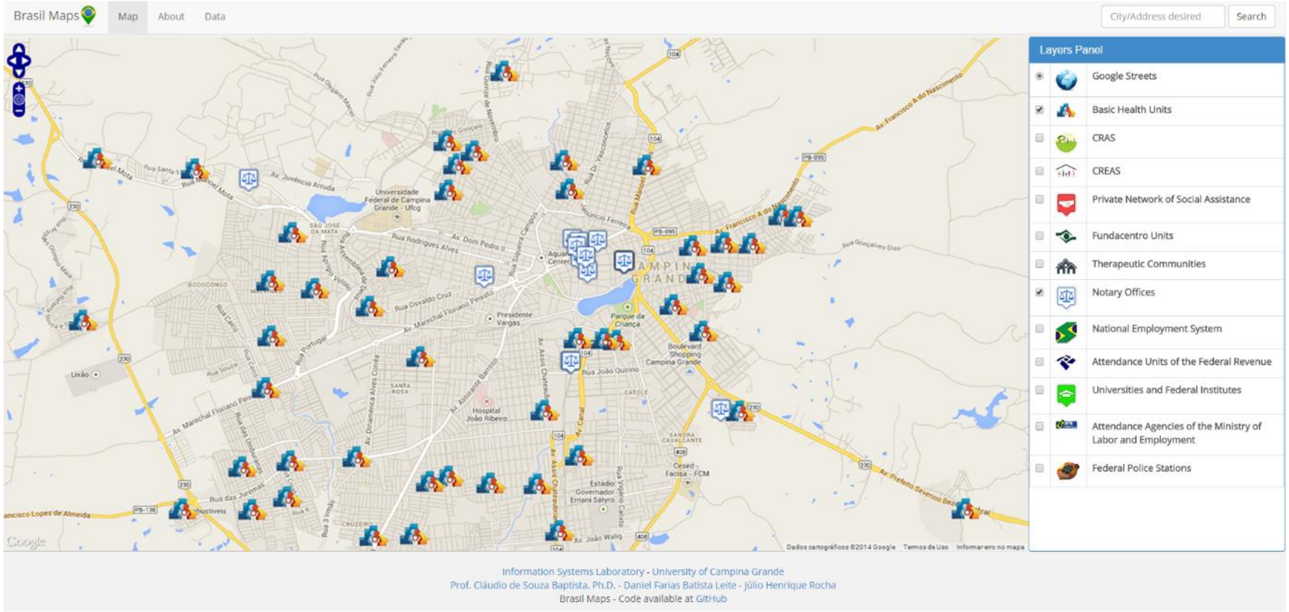
The Business Layer encompasses the *GeoServer* open source map server framework. *Geoserver* provides *Webmapping* development services. From a connection with the spatial database, *GeoServer* can produce a set of thematic maps which, in turn, will be consumed by the Application Layer from a *Web Map Service* (WMS) request. The geocoding module, that contains the geocoding service responsible for obtaining the geographic coordinates, is also located in the Business Layer.

Finally, the Data Layer is composed of an object relational database with support to spatial information (PostgreSQL + PostGIS). This layer is responsible for persisting and managing all the application's data, dealing with SQL queries and spatial functions.

B. *Functionalities*

The spatial location of all twelve public equipments available in the application needs to be presented in a map. In this context, ***BrasilMaps*** contains a map viewer (*WebMapping*), with the zooming and panning functionalities. The map displayed on the application was built using the *OpenLayers* library, that provides the visualization of spatial data.

Each public equipment is displayed in the map as a layer. This way, ***BrasilMaps*** enables overlaying the information, in which different equipment layers may be combined in a map. Figure 6 presents ***BrasilMaps*** main interface, which is launched with the map centered on the user's location, with the layer "Notary Offices" and "Basic Health Units" enabled. On the right side of the image is shown the *LayerSwitcher* component, responsible for managing the layers.



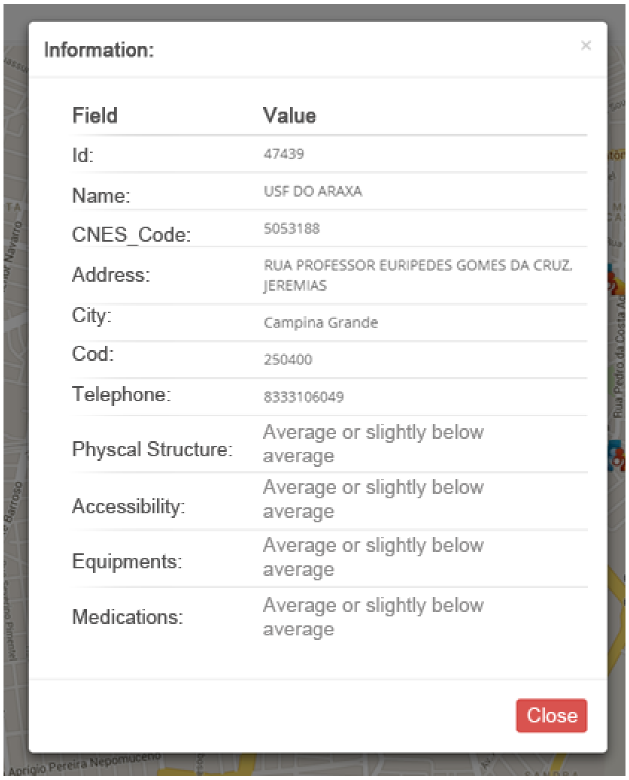
● Google Streets
☑ Basic Health Units
□ CRAS
□ CREAS
□ Private Network of Social Assistance
□ Fundacentro Units
□ Therapeutic Communities
☑ Notary Offices
□ National Employment System
□ Attendance Units of the Federal Revenue
□ Universities and Federal Institutes
□ Attendance Agencies of the Ministry of Labor and Employment
□ Federal Police Stations

Information Systems Laboratory - University of Campina Grande
Prof. Cláudio de Souza Baptista, Ph.D. - Daniel Farias Batista Leite - Júlio Henrique Rocha
Brasil Maps - Code available at GitHub

Figure 6. ***BrasilMaps*** main interface with two layers enabled.

Through user's location ***BrasilMaps*** assists the user in finding services of his interest. For example, let us consider a user that has been in an accident at his home and needs fast medical care. Hence, ***BrasilMaps*** can, using user location, point the Basic Health Units closer to him, facilitating the search for medical help and probably avoiding greater damages due to delayed medical support.

It is vital for a system that embraces and displays public services to provide the details for each of these services. Thus, ***BrasilMaps*** allows users to view the details of each entity displayed on the map using an information window. To use this functionality, the user must click on the desired public equipment, which will cause a window to popup displaying all the basic and specific information of the equipment, such as phone number, physical structure and others. This capability enables the refinement of the desired searches. In the example given above, the user might analyze whether the chosen Basic Health Unit would meet his needs. Figure 7 presents the information window of a Basic Health Unit.

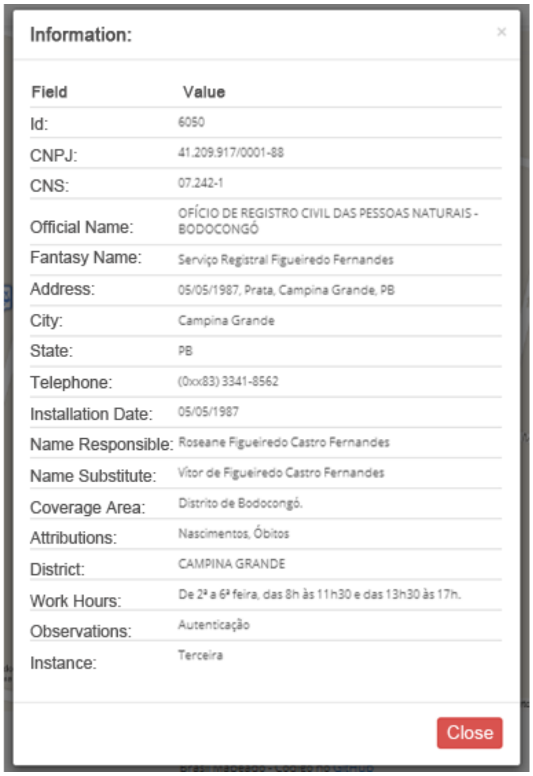


Information:

Field	Value
Id:	47439
Name:	USF DO ARAXA
CNES_Code:	5053188
Address:	RUA PROFESSOR EURIPEDES GOMES DA CRUZ. JEREMIAS
City:	Campina Grande
Cod:	250400
Telephone:	8333106049
Physcal Structure:	Average or slightly below average
Accessibility:	Average or slightly below average
Equipments:	Average or slightly below average
Medications:	Average or slightly below average

Figure 7. Popup with the details of a Basic Health Unit

With the help of the information window, a user that wishes to register a recently acquired property can check if the closest notary office to his current location may perform such task, as shown in Figure 8.



Information:

Field	Value
Id:	6050
CNPJ:	41.209.917/0001-88
CNS:	07.242-1
Official Name:	OFÍCIO DE REGISTRO CIVIL DAS PESSOAS NATURAIS - BODOCONGÓ
Fantasy Name:	Serviço Registral Figueiredo Fernandes
Address:	05/05/1987, Prata, Campina Grande, PB
City:	Campina Grande
State:	PB
Telephone:	(0xx83) 3341-8562
Installation Date:	05/05/1987
Name Responsible:	Roseane Figueiredo Castro Fernandes
Name Substitute:	Vítor de Figueiredo Castro Fernandes
Coverage Area:	Distrito de Bodocongó.
Attributions:	Nascimentos, Óbitos
District:	CAMPINA GRANDE
Work Hours:	De 2ª a 6ª feira, das 8h às 11h30 e das 13h30 às 17h.
Observations:	Autenticação
Instance:	Terceira

Figure 8. Popup with the details of a notary office.

Informing the user about the specific services offered by government services close to him promotes knowledge and avoids unnecessary displacement to obtain such information.

V. USABILITY ASSESSMENT

In order to assess the usability of the proposed system, we performed an evaluation consisting of analyzing its conformity to the standards defined in chapter 14 of ISO 9241 - the standard from the International Organization for Standardization that covers ergonomics of human-computer interaction. The recommendations described in such chapter of ISO 9241 suggest how the user interaction with the system might occur, considering specifically the menu structure and presentation, navigation, option selection and execution. So, such recommendations were evaluated in ***BrasilMaps*** and checked if they were successfully adopted. The results are shown in Table 1.

TABLE I. RESULTS OF THE STANDARD CONFORMITY EVALUATION

Standard	**#Sar**	**#Ar**	**AR (%)**
ISO 9241 Part 14	31	37	83,78

Where #Sar is the number of successfully adhered recommendations, #Ar is the number of applicable recommendations and AR is the Adherence Rate (AR = (Sar / Ar) * 100).

Therefore, we can observe that the adherence rate is over 80% for the analyzed chapter, meaning that ***BrasilMaps*** is in very acceptable conditions of use.

VI. Conclusions and Future Work

This paper presented an overview of ***BrasilMaps***, a GIS application that aims to assist the citizens in finding the desired government services closer to their location. As future work, we plan to incorporate crowdsourcing techniques combined with VGI (Voluntary Geographic Information) for feeding ***BrasilMaps*** with information obtained from social networks.

With the goal of expanding the user experience with the system, in the future ***BrasilMaps*** will support spatial searches operations and will allow the generation and visualization of routes between the user location and the desired public service. Lastly, the integration with other government databases, whether national or international, is of high importance. However, the adoption of international databases will create new system evolution possibilities and challenges that will be explored in future works.

References

[1] The Annotated 8 Principles of Open Government Data, available at http://opengovdata.org/ [accessed: 2015-01-05].

[2] The Three Laws of Open Government Data, available at http://eaves.ca/2009/09/30/three-law-of-open-government-data/ [accessed: 2015-01-05].

[3] R. Matheus, M. M. Ribeiro, and J. C. Vaz, “New perspectives for electronic government in Brazil: the adoption of open government data in national and subnational governments of Brazil”,in Proc. of the 6th International Conference on Theory and Practice of Electronic Governance - ICEGOV, October 2012,pp.22-29.

[4] BrasilMaps, available at http://150.165.75.163/brasil-mapeado [accessed: 2014-10-25].

[5] Mootiro Maps, available at http://maps.mootiro.org/project/307/map [accessed: 2014-12-21]

[6] PAC II, available at http://pac-info.herokuapp.com/map.html [accessed: 2014-12-21]

[7] L. Pedroso, A. Tanaka, and C. Cappelli, “The Brazilian law of access to information and the technological challenges of open government data”, Brazilian Symposium of Information System,May 2013, pp. 523 – 528.

[8] R. P. Lourenço and l. Serra, “An Online Transparency for Accountability Maturity Model”, in Proc. of the 13th IFIP WG 8.5 International Conference- EGOV, September 2014, pp. 35-46.

[9] K. Breitman et al. “Open Government Data in Brazil”, IEEE Intelligent Systems, vol. 27, no. 3, November 2012, pp. 45-49.

[10] H. Xu and I. Zheng, “Open government data: from users’ perspectives”, in Proc. of the 7th International Conference on Theory and Practice of Electronic Governance - ICEGOV, October 2013, pp. 366-67.

[11] A. S. Correa, P. L. Correa, and F. S. C. Silva, “Transparency Portals versus Open Government Data. An Assessment of Openness in Brazilian Municipalities”, in Proc. of the 15th Annual International Conference on Digital Government Research – DG.O, June 2014, pp. 178-85.

[12] J. S. Erickson, A. Viswanathan, J. Shinavier, Yongmei Shi and J. A. Hendler, “Open Government Data: A Data Analytics Approach”, IEEE Intelligent Systems, vol. 28, no. 5, March 2013, pp. 19-23.

[13] F. Artigas, “Spatial analytics for open government data”, in Proc. of the 15th Annual International Conference on Digital Government Research – DG.O, June 2014, pp. 357-58.

[14] F. Artigas and S. A. Chun, “Visual analytics for open government data”, in Proc. of the 14th Annual International Conference on Digital Government Research – DG.O, June 2013, pp. 298-99.

[15] A. Graves and J. Hendler, “Visualization tools for open government data”, in Proc. of the 14th Annual International Conference on Digital Government Research – DG.O, June 2013, pp. 136-45.

Data Concretization States as Metadata of Parameterized Regional Futures in a WebSDSS Development Context

Rico Vogel and Marco Neubert and Axel Sauer

Leibniz Institute of Ecological Urban and Regional Development (IOER)
Dresden, Germany
Email: {r.vogel, m.neubert, a.sauer}@ioer.de

***Abstract*—The development of a web-based spatial decision support system (WebSDSS) aims at the transfer of knowledge that arises from the operationalization of a scenario approach that focuses on so-called 'parameterized regional futures' (PRFs). Because the operationalization of the PRF approach is an ongoing process, a gap exists between data required and currently available. To overcome this gap and to allow specific software testing, test data sets can substitute expected PRF data. Based on the nature of spatial and non-spatial data and its role in different WebSDSS development phases, this article proposes an approach to distinguish concretization states as metadata of original PRF data and test data counterparts. This contribution accordingly addresses topics like modeling or managing spatial data, geospatial domain applications, and digital cartography.**

Keywords–Software development; WebSDSS development; Metadata information; Data concretization states; Test data; Parameterized regional futures; Scenario approach.

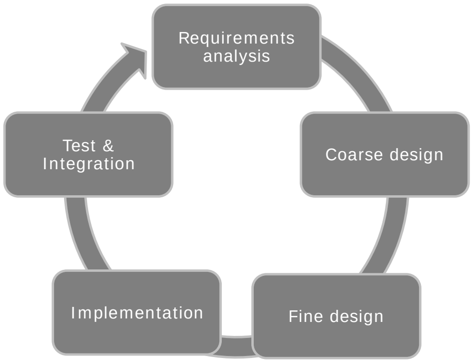


Figure 1. Iterative incremental development and method steps (modified according to [2]).

I. Introduction

Currently, a web-based spatial decision support system (WebSDSS), characterized by Rinner [1] for example, is being developed by Vogel [2] and the related research group. This tool aims at enabling the immediate transfer of knowledge, which arises from the operationalization of a scenario approach of Schanze and Sauer [3] that focuses on so called 'parameterized regional futures' (PRFs, see also [4]). This software accordingly provides steered access to the complex pool of spatial and non-spatial PRF data by using web-based geographical information system (WebGIS) technology, which facilitates thematical, temporal, and spatial selection, preparation, and presentation of such data via the Internet.

Within a PRF, a *future* is the core component composed of one specific *scenario* and one specific *strategic alternative*. A *scenario* consists of a set of *projections* derived from a narrative storyline addressing climate change, demographic change, technological change, economic change, as well as land-use change (e.g., increasing temperatures, aging society). A *strategic alternative* comprises different *intervention options* that are single and partly site-specific measures [4]. Building a dike or improving flood resilience of buildings are examples for such options. Both, *projections* and *intervention options* are defined by further sub components. A formalization of the PRF approach capturing the component hierarchy has been prepared by [5].

In the case it serves as the leaf of the PRF hierarchy, each sub-component is commonly expected to be represented by a spatial data set, which may be derived according to [6]. Dependent on the number of *futures* to be operationalized, a number of slightly varying data sets is expected for each theme (leaf sub-component). In the case of the PRFs, about a dozen of *futures* are operationalized [3].

This article proposes an approach to distinguish concretization states of original PRF data and types of test data. First, the role of spatial and non-spatial data in the WebSDSS development process is explained (Section II). Rainardi [7] (pp. 477–489) emphasize the importance of test data in data warehouse development, and [8] exemplifies its value in conformance testing in the context of standardization. Afterwards, the nature of PRF data and associated test data is captured in Section III. Relevant data concretization states are identified in Section IV. Test data of such states can support different phases of WebSDSS development (e.g., [9], p. 299). Therefore, in user tests, data concretization states are important metadata information provided by web services like Web Mapping Services (WMSs). Finally, Section V concludes this article.

II. WebSDSS Development and Data Availability

The WebSDSS development intended by Vogel [2] follows an iterative-incremental approach. According to Kleuker [11] (pp. 30 f.) in particular the *requirements analysis*, *coarse design*, *fine design*, *implementation*, as well as *test & integration* phases may be distinguished. As shown in Figure 1, these phases are repeated; and, according to Rumbaugh, Jacobson, and Booch [12] (p. 319), each iteration results with an executable system that can be executed, tested, and debugged. Nevertheless, each of the phases may have certain requirements with regard to the availability of PRF data. While some like *coarse design* may be proceeded based on simple theoretical

ISBN: 978-1-61208-383-4

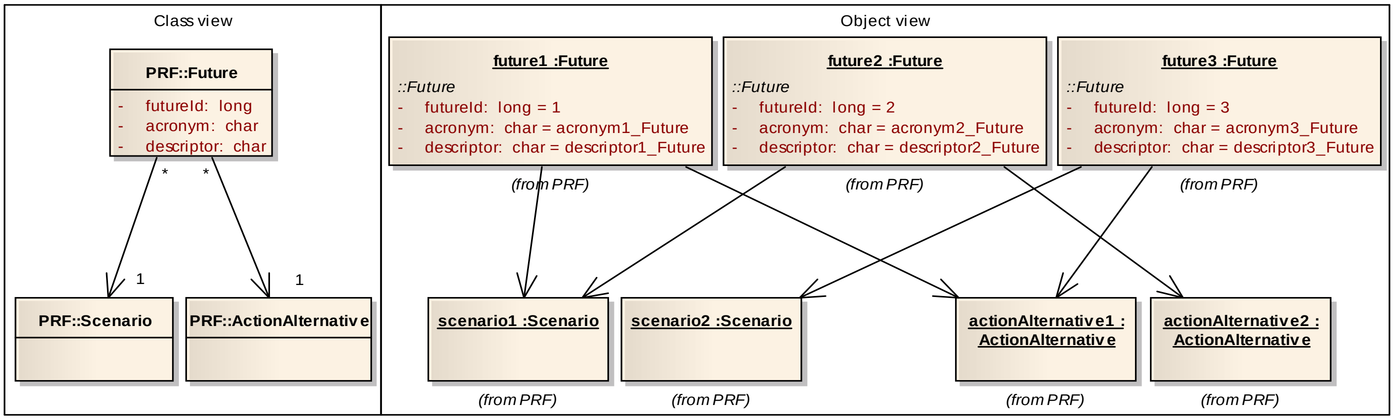


Figure 2. Concept of the *Future* component (Class view) and instances of the *abstract* data conretization state (Object view) used by *software developers* in early iterations of (conventional) software development (inspired by [10]).

examples, others like *test & integration* may require original data sets or at least such that approximate them.

Since PRF operationalization is an ongoing process that captures recent (e.g., [13] contributes to the 'Integrated Regional Climate Adaptation Programme for the Model Region Dresden', REGKLAM, research project) and current research (e.g., [6] contributes to the 'Vulnerability Study Saxony' (VusS) and to the 'Cross-Process Analysis of Vulnerabilities and Risks of Urban Regions with Respect to Climatic Factors and their Changes – Conceptualisation and Modelling' (RegioRisk) research projects), the complete set of multi-dimensional data (e.g., [14]) is not yet available for all PRFs to be operationalized. Rather, certain PRF aspects have already been analysed (e.g., [13]) or are work in progress (e.g., [6]).

Besides the intended database of PRF data, for the WebSDSS development there are several reasons to build up a test database (see Figure 2). For example, in early iterations simple use case examples are sufficient, while during *test & integration*, test data are required that at least embraces a sub area by a selection of PRF themes. Further data sets improve later iterations to *design* and *test* the system behavior for extreme data values, which are not expected from causal PRF operationalization (e.g., increasing average temperatures by 20 K in 5 years).

To fill the gap between the availability of real PRF data and test data (see Figure 3), an approach needs to be developed that takes into account the data requirements for each development phase as well as the opportunities to provide sufficient phase-related test data. Therefore, in a first step, important criteria have to be derived that enable the differentiation of (test) data requirements. Afterwards, valid data concretization states are suggested that seem most important in the context of PRF-specific WebSDSS development.

III. Capturing the Nature of PRF Data and Associated Test Data

From a development viewpoint, this section exemplifies the excerpt of important criteria, which enable the capturing of the nature of PRF data. These findings will be used to discuss proposed categories of *futures* in the next section. To capture the criteria, social, spatial, development, and scientific scopes

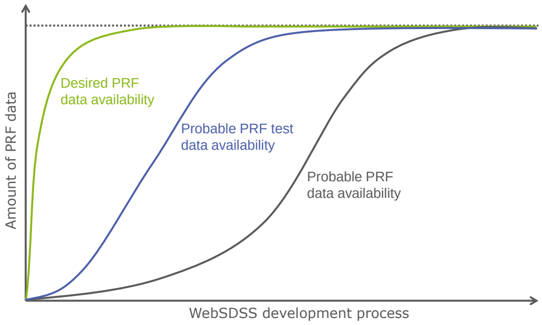


Figure 3. Schema of gaps between desired and probable availability of PRF data in WebSDSS development and supporting role of test data.

were differentiated. These metadata may be the source for characterizing different data concretization states.

From a social perspective, the main addressees of certain PRF data are distinguished. *Practitioners* and *planners* may be typical addressees of real PRF data. In contrast, test data are applicable in discussions during *requirements analysis* including further stakeholders such as the *scientists*, which developed the PRF approach (i.e., [3]).

Taking the spatial dimension into account leads to a distinction of *spatial* and *non-spatial* aspects of PRF data. Furthermore, the scale of a pilot region compared with the whole study region is differentiated. In particular, selecting *micro*, *meso*, or *macro* levels influences the amount of data that has to be processed during the *test & integration* phase for example.

During the technical WebSDSS development, it may be of importance to which extent the data sets cover the study area or if the data sets have limited spatial coverage (*complete*, *incomplete*). Additionally, data protection is an important issue. *High*, *moderate*, *low*, as well as *zero* protection levels can be distinguished.

Finally, scientific aspects can be taken into consideration. In particular, the scientific origin is of importance. For PRFs, the projects REGKLAM, VusS, and RegioRisk are itemized leading to *regklam*, *vuss*, and *regiorisk* identifiers. The data source can be differentiated in terms of scientific products such as internal documents (*report*), publications (*article*, *proceeding*) or data sets (*dataset*, *map*).

IV. Data Concretization States

Based on the criteria identified in Section III, certain concretization states for data sets of the PRF approach are determined, which have importance for the WebSDSS development. Against the backdrop of the characterization of PRF data and test data in the previous section, four different data concretization states can be distinguished: *abstract*, *arbitrary*, *mockup*, as well as *genuine* (see Figure 4). These can shortly be described as follows:

- *Abstract data sets* are used to outline concepts of components and relationships among components in a simple manner. Often, simple characters are used to mark objects (e.g., by *a*, *b*, *A*, *B*) or such characters are combined with the type name of an object, for example *aTown* and *bTown* are used to distinguish two towns. In certain development phases, abstract data may be sufficient; in particular in *fine design* realized by a *software developer* simple examples are needed. In contrast, *test & integration* involving local and regional *planners* and *scientists* will be impractical if use case examples are merely based on *abstract data*.
- *Arbitrary data sets* are likewise valuable especially if spatial data is required in early development phases. Comparable with 'lorem ipsum' text used in typography, *arbitrary* data sets may be used as placeholders, for example to test map components by consuming map services. Data of this concretization state will primarily be helpful for the *software developer*. The current prototypical implementation [16], for example, simply associates accessed ArcGIS representational state transfer (REST) services, also used by Strode [17], to certain PRF component instances, which itself are merely of *abstract* state (e.g., sub-components of a *scenarioC* and a *strategicAlternativeB*).
- Unlike data of the previous concretization states, *mockup* data sets approximate (e.g., Benenson and Torrens [18], p. 16) or even consider the conditions of the study region. Although, they are still fictional, *mockup* (or dummy/ synthetic) data sets are designed to approximate *genuine* ones (comparable with 'potemkin villages'), since they are often the result of the same or similar models like used for the *genuine* PRF data. Accordingly, *mockup* data are assumed to be close to reality which therefore enable nearly-realistic software *test & integration*, involving stakeholders such as local and regional *planners*. Furthermore, *mockup* data sets may be created (e.g., simulated according to [18]) to offer (extreme) data values that are not expected to appear within *genuine* data sets.
- Finally, *genuine* data sets are the intended results gathered from the realization of the PRF approach. From the development point of view, data of this concretization state is valuable in each of the development phases. But in the subsequent production phase, merely *genuine* data sets are valid to build up the WebSDSS data component.

V. Conclusion

The article demonstrated an approach for capturing data concretization states of PRF data. Within WebSDSS development, these states are influencing each phase. For example, during *implementation* and associated unit *testing* by a *developer*, the state of PRF data, consumed by a web service may be registered by a logging mechanism. Of equal importance is the fact that such metadata information should be visualized by a graphical user interface (GUI) during *test & integration* to inform involved *planners* and other stakeholders about the concretization states of the data displayed.

Therefore, it is highly recommended to mark the data sets by their concretization state. Especially with a huge amount of data, this marking allows for the fast investigation of the required degree of replacement of test data by *genuine* one. Table I summarizes concretization states of PRF data, which at least have to be available in early software development iterations (x) or even in the production phase. Data sets may be swapped in the course of time, for example, by replacement with more concrete ones (o). To ease the access on such state information, metadata extensions or profiles according to [19] may be appropriate. To avoid misinterpretation of data content, induced by lacked information of the states of data consumed, especially in the case of mockup data, it is strongly recommended to automatically add the data concretization state information during data set production, or during the creation of related web services to the data and web service metadata.

TABLE I. Data concretization states in WebSDSS development and production (x – minimum requirement / o – aimed substitutions in subsequent iterations).

Development phase	*abstract*	*arbitrary*	*mockup*	*genuine*
Requirements analysis	x	o	o	o
Coarse design	x	o	o	o
Fine design	x	o	o	o
Implementation		x	o	o
Test & integration			x	o
Production				x

Nevertheless, considering aggregates on upper concretization levels requires more specific distinctions among aggregates, because the concretization states of an aggregate depends on the ones of its sub components. In upcoming work, the presented approach will be refined and adapted for such aggregates. In this context, the taxonomy defined by the domain model of [5] may be used to identify the aggregates and its members. As already mentioned, each *future* references exactly one *scenario* and one *strategic alternative*. Accordingly, its concretization state results from those of both sub components. For example, while a *scenarioC* might be marked with the *arbitrary* state, the state of a *strategicAlternativeB* is of type *mockup*. For such cases, intermediate operationalization states are currently under development. Vogel [20] proposes appropriate alternatives such as *abstract prevailed*, *arbitrary prevailed*, as well as *mockup prevailed*.

The concept presented in this article, is also applicable to data of further scenario approaches. Beyond the scope of WebSDSS development, it can also improve conventional (non-spatial) software development. Even in offline development like spatial analysis, data concretization states can be attached to metadata of geoprocessing workflow results, for

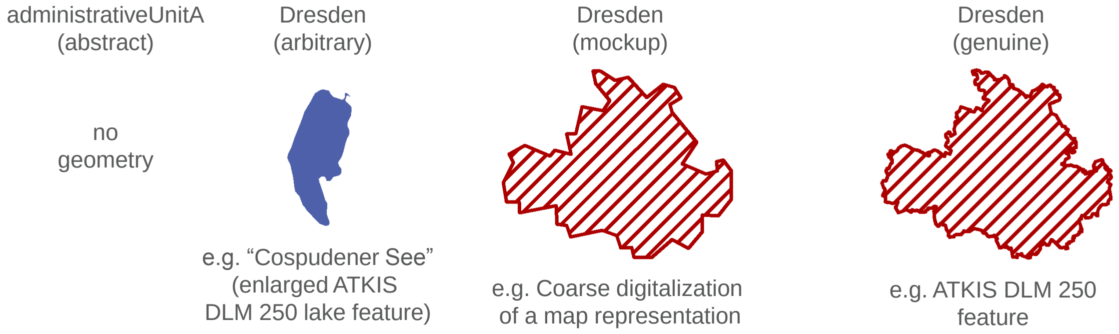


Figure 4. Data concretization states at the example of possible instances of an *AdministrativeUnit* feature type according to INSPIRE [15].

example to improve interpretation of exchanged data by project members.

ACKNOWLEDGMENT

The research described in this paper emerged from developments in the 'Design, Technologies and Usability of web-based Spatial Decision Support Systems for regional assessment of climate change impacts and adaptation – for the Model Region of Dresden' (ZENON; sCENarios ONline), a research project funded by the Leibniz Institute of Ecological Urban and Regional Development (IOER), Dresden, Germany. Furthermore, the authors thank colleague Robert Hecht and several anonymous reviewers for their advice and time.

REFERENCES

[1] C. Rinner, "Web-based Spatial Decision Support: Status and Research Directions," Journal of Geographic Information and Decision Analysis (GIDA), vol. 7, no. 1, 2003, pp. 14–31, URL: http://www.ryerson.ca/∼crinner/pubs/Rinner_published_GIDA-01-2003.pdf [accessed: 2015.01.08].

[2] R. Vogel, "ZENON – Entwicklung eines WebSDSS zur Abschätzung der Folgen des Klimawandels und des gesellschaftlichen Wandels [ZENON – Development of a WebSDSS for assessment of impacts of climate change and societal change]," in Angewandte Geoinformatik 2013: Beiträge zum 25. AGIT-Symposium Salzburg, J. Strobl, T. Blaschke, G. Griesebner, and B. Zagel, Eds. Berlin, Offenbach: AGIT Salzburg, 2013, pp. 227–232, URL: http://gispoint.de/fileadmin/user_upload/paper_gis_open/537533082.pdf [accessed: 2015.01.08].

[3] J. Schanze and A. Sauer, "Dokumentation des fragestellungs- und gebietsspezifischen Szenarioansatzes [Documentation of the topic-related and region-specific scenario approach]," Leibniz-Institut für ökologische Raumentwicklung, Dresden, REGKLAM Ergebnisbericht, Produkt 2.4a, Version: 1, Status: Konzept, unpublished, 2011.

[4] J. Schanze, J. Trümper, C. Burmeister, D. Pavlik, and I. Kruhlov, "A methodology for dealing with regional change in integrated water resources management," Environmental Earth Sciences, vol. 65, 2012, pp. 1405–1414.

[5] R. Vogel, A. Sauer, and J. Schanze, "Object-Oriented Domain Modeling of a Regional Climate Change Scenario Approach as a Central WebSDSS Development Artifact," N. N., in preparation.

[6] M. Neubert and R. Vogel, "Integrierte räumliche Schadensanalysen zum Aufbau der Datenhaltungskomponente eines webbasierten Entscheidungsunterstützungssystems [Integrated spatial damage analysis for establishing the database component of a web-based decision support system]," in Angewandte Geoinformatik 2014: Beiträge zum 26. AGIT-Symposium Salzburg, J. Strobl, T. Blaschke, G. Griesebner, and B. Zagel, Eds. Berlin, Offenbach: AGIT Salzburg, 2014, pp. 375–380, URL: http://gispoint.de/fileadmin/user_upload/paper_gis_open/537543055.pdf [accessed: 2015.01.08].

[7] V. Rainardi, Building a Data Warehouse With Examples in SQL Server. Berkeley: Apress, 2008.

[8] OGC, "OpenGIS® Implementation Specification for Geographic information – Simple feature access – Part 2: SQL option," Open Geospatial Consortium, OpenGIS® Implementation Specification OGC 06-104r4, Version 1.2.1, 2010, URL: http://www.opengeospatial.org/standards/sfs [accessed: 2015.01.08].

[9] D. Haywood, Domain-Driven Design: Using Naked Objects. Raleigh, Dallas: The Pragmatic Bookshelf, 2009.

[10] R. Vogel, "Testanwendung eines technologieintegrierenden WebSDSS-Prototyps – Dokumentation des 2. Tests [Test application of a technology integrating prototype of a WebSDSS – Documentation of the 2nd test]," ZENON AP5 project report, unpublished, 2014.

[11] S. Kleuker, Grundkurs Software-Engineering mit UML: Der pragmatische Weg zu erfolgreichen Softwareprojekten [Basic course software engineering with UML. The pragmatic way to successful software projects], 2nd ed., ser. Studium. Wiesbaden: Vieweg + Teubner, 2011.

[12] J. Rumbaugh, I. Jacobson, and G. Booch, The Unified Modeling Language Reference Manual. Reading, Massachusetts: Addison-Wesley, 1999.

[13] O. Kretschmer, "GIS-basierte Projektionen der Siedlungsflächenentwicklung mittels multikriterieller Bewertungsfaktoren und zellulärem Automatenmodell DINAMICA [GIS-based projections of residential area dynamics with multi-criteria evaluation factors and the cellular automata model DINAMICA]," Diploma thesis, IOER Dresden and Philipps-Universität Marburg, unpublished, 2012.

[14] A. Yeung and G. Hall, Spatial Database Systems: Design, Implementation and Project Management, ser. The GeoJournal Library 87. Dordrecht: Springer-Verlag, 2007.

[15] INSPIRE, "Inspire data specification on administrative units – guidelines," INSPIRE Thematic Working Group Administrative units, Data Specification D2.8.I.4, 2009-09-07 2009, URL: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_AU_v3.0.pdf [accessed: 2015.01.08].

[16] R. Vogel, "Prototypische Technologieintegration für ein WebSDSS zur Folgenabschätzung des Klimawandels und des regionalen Wandels [Prototypical Technology Integration for a WebSDSS for Assessment of Impacts of Climate Change and Regional Change]," FB R-Kolloquium, presentation, 2012.08.08, IOER Dresden, 2012.

[17] G. Strode, "The GIS Behind iMapInvasives: The "Open Source Sandwich"," in Online Maps with APIs and WebServices, ser. Lecture Notes in Geoinformation and Cartography, M. Peterson, Ed. Berlin, Heidelberg: Springer, 2012, ch. 6, pp. 73–90.

[18] I. Benenson and P. M. Torrens, Geosimulation: Automata-based Modelling of Urban Phenomena. Southern Gate, Chichester, West Sussex: John Wiley & Sons, 2004.

[19] ISO 19115, "Geographic information – Metadata," International Organization for Standardization, ISO 19115:2005-05, 2005.

[20] R. Vogel, "Zukünftedaten im Kontext der Entwicklung eines WebSDSS zur Analyse von Szenarien des Klimawandels und des regionalen Wandels [Future data in the context of the development of a WebSDSS for analysing scenarios of climate change and societal change]," ZENON AP4 & AP5 project report, unpublished, 2014.

A new *n*D Temporal Geodata Management Approach using GML

Martin Hoppen, Michael Schluse, Juergen Rossmann

Institute for Man-Machine Interaction
RWTH Aachen University
Aachen, Germany
Email: {hoppen,schluse,rossmann}@mmi.rwth-aachen.de

Christoph Averdung

CPA Geo-Information
Siegburg, Germany
Email: averdung@supportgis.de

***Abstract*—Geodata represents the state of a real world phenomenon at some point in time. Multiple, differently interpreted reference times can be associated to the same datum yielding an n-time-dimensional (nD) problem. This reference time, however, is often ignored, reduced or managed ”manually” on application level. We propose a new approach for nD temporal geodata management encapsulating and hiding this complexity. The Geography Markup Language (GML) and a transactional Web Feature Service (WFS-T) are used for standard compliant data exchange. Its feasibility is proven by a reference implementation and applications. Thus, the main achievement of this paper is the presentation of a new nD temporal data management approach using time point representation and its application to GML-based forestry applications.**

***Keywords*–Geodata Management; Temporal Data Management; Temporal Databases; GML; WFS-T.**

I. Introduction

In geodata, the concept of time is pervasive. Geodata represents the state of a real world phenomenon at one or more points in time. Orthogonally, the data itself may be captured, manually revised, or updated at different times. There may also be some point in time where the geodata becomes effective, i.e., valid in a formal sense. Finally, when incorporating geodata into simulation models, simulation time can be considered, as well. These different interpretations of time values are called *time dimensions* [1].

Thus, when dealing with geodata, its time references can only be omitted in the simplest cases. An appropriate temporal geodata management is therefore advisable. However, managing temporal references may be a complex, tedious and error-prone task – in particular when performed ”manually” on application level. Instead, one should prefer a transparent data management layer, encapsulating and hiding this complexity. Another advantage of such an application-independent approach is that the same management layer can be reused for different applications. Regarding the aforementioned diversity of time dimensions, the underlying temporal modeling concept should furthermore be sufficiently universal.

In previous publications, we presented the fundamentals for using temporal data management in 3D simulation applications [2]. Subsequently, we mentioned the idea of an n-time-dimensional versioning concept for geodata [3]. In this paper, we introduce a detailed mathematical specification of this concept and demonstrate its feasibility by means of a reference implementation on geodatabase and on client application level. Furthermore, we show how this higher-order temporal metadata is marshaled into a GML representation for data exchange – in particular based on WFS-T. Finally, the approach is used as a basis for distributed parallel data management.

The rest of the paper is structured as follows. In Section II, we give an overview of the state of the art. In the following section (Section III), the new concept for *n*D-versioning is introduced. Its realization in a geodatabase system and integration into a 3D simulation system is presented in Sections IV–V. The approach is applied to a distributed parallel data management scenario in Section VI presenting a corresponding application in Section VII. We conclude the paper in Section VIII.

II. State of the Art

A. Temporal Databases

As described in [4][5], in general, a temporal database is any database containing some temporal information. This can be explicitly modeled within the scope of the application schema, e.g., a date field ”start of employment” for an employee’s record. Here, application programs have to manage temporal data on their own, which can be a complex task for sophisticated scenarios. To cope with this complexity, different concepts have been developed generalizing the problem.

Furthermore, a time value must be interpreted. Such interpretations are called time dimensions [1]. The time associated with an event can represent its occurrence in the real world. This time dimension is called *valid time*, the corresponding database is called a *valid time database*. When time values refer to the point in time when the values were stored in the database, it is called *transaction time*, the corresponding database *transaction time database*. A database combining both time dimensions is a bitemporal database. Besides these two most common time dimensions there can be more interpretations.

When using valid time, for each operation (read or write) the user has to supplement the point in real time corresponding to the event. This also allows to specify future or past events leading to so-called *proactive* or *retroactive updates*. When using transaction time, however, timestamps for writing operations are automatically derived from the database’s system clock; proactive or retroactive updates are impossible. For reading operations omitting to specify a transaction time, the current (i.e., last written) value is returned. This can also be specified as a transaction time of *cur* (current). For other points in time, the corresponding historic value is returned.

When updating a data item in a temporal database, its previous state (i.e., version) is not lost. Rather, a new version with the new value is inserted and the previous version is closed becoming a so-called history version. Thus, every data item is represented by a set of $n{>}0$ versions. The state of a temporal database at a certain point in time (or a tuple thereof) is called a snapshot. Likewise, non-temporal databases are called snapshot databases.

A usual approach is to store start as well as end timestamps for each version and time dimension. While every version must have start timestamps, end times are only set to close a version. Deleting a data item is only performed logically by closing its current version. Inserting a new data item creates an initial version. Following this standard approach, in bitemporal databases, an update even creates two new versions. By updating a data item, its current version is closed concerning valid time by setting an end timestamp. The timestamp's update itself however leads to a new version regarding transaction time but still holding the previous value of the data item. The second version is created with the actual new value and current timestamps.

In relational databases, this concept can be implemented by adding start and end timestamp attributes for each time dimension to every relation. This has the drawback of redundancy as changing a single attribute value of a tuple creates a new version duplicating all other attribute values. This can, for example, be reduced by dividing the relation into subrelations. In object-oriented or object-relational databases, a common approach is attribute-based versioning. Here, usually, an attribute version is stored as a tuple of start and end time for each time dimension and the corresponding value. Thus, for each of an object's attributes, a list of such tuples is stored. In addition, the object itself maintains so-called *lifespan* temporal attributes for each time dimension. They represent its overall lifetime from creation to deletion.

Following [6], the aforementioned approach of storing start as well as end times for each version is called an *interval representation*. In contrast, storing only start times is called a *time point representation*. Here, subsequent versions' start times close preceding versions. Consequently, each version has a certain *scope* comprising those points in time the version's value is "visible."

An extensive overview of existing spatio-temporal database models is given in [7]. While about $2/3$ of the systems are bitemporal, in contrast to the proposed approach, none provides three (or more) time dimensions. Furthermore, only about $1/3$ of the proposed models have been implemented and about $1/3$ lack a formal representation. Other, more recent approaches like [8] do integrate more dimensions (scale) but reduce time to one dimension. Finally, a very recent publication [9] presents a substantial online bibliography on various aspects of temporal GIS, which has yet to be examined.

B. Temporal Data Management Standards

The International Organization for Standardization (ISO) norm 19100 [10] series of geographic information standards in combination with the specifications of the Open Geospatial Consortium (OGC) [11] provide the basis for designing the structure of time-related geospatial data. For a standard-based design of geospatial data structures with a focus on database-driven data management, the following ISO standards are important:

- ISO 19119: Geographic information - Services
- ISO 19109: ... - Rules for application schema
- ISO 19107: ... - Spatial schema
- ISO 19111: ... - Spatial referencing by coordinates
- ISO 19125: ... - Simple feature access
- ISO 19136: ... - Geography Markup Language (GML)

ISO 19108 adds specifications for the description of time-related issues. This is the basis for extending the usually 3-dimensional structure of geospatial reference systems by temporal properties. For that purpose, ISO 19108 provides rules for the schema design of time as an absolute (point of time, duration) or a chronological (before, after, in-between) specification of time.

The geospatial data's reference time can be modeled in different ways, depending on the technical requirements of the given task. It can either be part of the data's schema itself (*explicit modeling*) or it can be a built-in property of the applications system architecture (*implicit modeling*). In both cases, ideally, the specifications of ISO 19108 are taken into account yielding an implementation that considers international standards.

III. *n*D-Versioning

As motivated in Section I, besides valid and transaction time (e.g., combined in a bitemporal database), even more time dimensions can be thought of. One of these is the effectivity of a change, i.e., the time it becomes effective or valid in a formal sense. For example, in data acquisition for a forestry application (compare Section VII), tree heights may be updated over a longer period of time. Each height value is associated with the appropriate valid timestamp (actual time of the measurement) and transaction timestamp (time of the database entry). The third timestamp represents the time a change becomes effective, e.g., the beginning of the next inventory period at the end of all measurements. This allows to decouple the description of real world phenomena with valid time from the effectivity of these values while modeling the more or less technical information about data storage with transaction time. While this three-time-dimensional scenario shall be used to motivate and explain the following specification, our concept can be generalized to arbitrary kinds and numbers of time dimensions.

For this higher-order temporal database, a new concept for evaluating time point representations was developed. As mentioned above, in this case, historic versions are only implicitly closed by other versions' timestamps. We show how this representation can be applied to all integrated time dimensions at the same time.

A. One and Two Time Dimensions

To begin with, this approach can also be applied for one or two time dimensions. In the former case, every update is associated with a single (scalar) timestamp t_i with associated version v_i. The interval $[t_i, t_j)$, where $t_j{>}t_i$ is the next timestamp, defines v_i's scope (with $t_j{=}\infty$ if t_i is the last timestamp). A retroactive update (e.g., in case of a valid time database) with timestamp $t_k, t_i{<}t_k{<}t_j$ splits the existing interval into $[t_i, t_k)$ and $[t_k, t_j)$.

In the bitemporal case (i.e., with the two time dimensions transaction and valid time), an update is associated with a timestamp tuple (t_i^T, t_i^V) for transaction time T and valid time V. Here, the two-time-dimensional scope of the associated version v_i initially (without any other existing versions) is a quarter-plane restricted by the two rays emerging from (t_i^T, t_i^V) parallel to and in positive direction of the respective axis (dotted area in Figure 1 left).

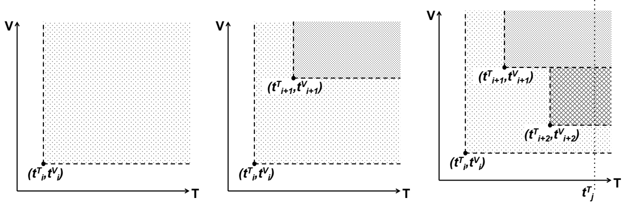


Figure 1. Bitemporal example of the evaluation schema for the higher-order temporal database.

Inserting a new version v_{i+1} with a timestamp tuple (t_{i+1}^T, t_{i+1}^V) with $t_i^T < t_{i+1}^T, t_i^V < t_{i+1}^V$ (Figure 1 center) reduces the scope of v_i to two infinite "tubes" while the new scope of v_{i+1} is again defined by a quarter-plane rooted at (t_{i+1}^T, t_{i+1}^V). Thus, value version v_{i+1} is only visible when considering a reference time greater or equal in both timestamp components. As transaction time does not allow retroactive updates, $t_i^T < t_{i+1}^T$ must always be valid. However, valid time allows for $t_i^V > t_{i+1}^V$. An example is shown in Figure 1 (right). A new version v_{i+2} with timestamp tuple (t_{i+2}^T, t_{i+2}^V) is inserted where $t_{i+1}^T < t_{i+2}^T, t_{i+1}^V > t_{i+2}^V$. While v_i's scope is similarly restricted as in the previous example, v_{i+2}'s scope stops at v_{i+1}'s. We call this behavior the "Golden Rule": The expansion of a limiting ray stops at preexisting other scopes. In the exemplary scenario, the expansion of the ray in V-direction (V-ray) shall stop. This allows retroactive updates with regard to valid time to "slide in between" existing values. That is, for any $t^T \geq t_{i+2}^T$, the scope of v_{i+2} is between v_i's and v_{i+1}'s scopes. As a corollary, T-rays do *not* stop at existing V-rays. In particular, this makes the scope of new versions independent of their insertion's order regarding valid time. In Figure 1 (right), the order of the valid timestamp components is irrelevant to the resulting partitioning along the V axis. Figure 2 gives an example for an alternate insertion order. The partitioning regarding valid time along any $t_j^T > t_{i+2}^T$ is identical to the prior case.

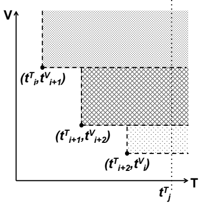


Figure 2. Inserting versions in an alternate order (i.e., different transaction times) yields the same partitioning along a $t_j^T > t_{i+2}^T$.

B. Three Time Dimensions

This evaluation schema can now be extended to the three time dimensions as mentioned above: transaction time T, valid time V, and effectivity time E. This leads to three-dimensional timestamps (t_i^T, t_i^V, t_i^E) for every version v_i. The resulting scope (to begin with, without interference of other timestamps) is a "cube" with infinite extent in direction of each axis. The simplest case can be compared to the 2D case in Figure 1 (center) and is shown in Figure 3: The scope for timestamp (t_i^T, t_i^V, t_i^E) is reduced to (up to) three "walls" by a new version with timestamp tuple $(t_{i+1}^T, t_{i+1}^V, t_{i+1}^E)$ if $t_{i+1}^T \geq t_i^T, t_{i+1}^V \geq t_i^V, t_{i+1}^E \geq t_i^E$. Note that – as for the 2D case – the actual scopes infinitely expand in (positive) axis directions.

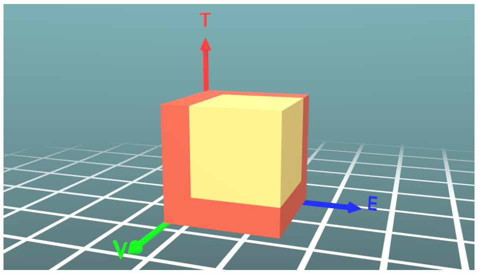


Figure 3. Intersection of i^{th} (orange) and $(i+1)^{th}$ (yellow) 3D time scope for $t_{i+1}^T \geq t_i^T, t_{i+1}^V \geq t_i^V, t_{i+1}^E \geq t_i^E$.

For other constellations, again, the "Golden Rule" resolves ambiguities. In the 3D case, the rays emerging from a timestamp tuple in each axis direction have to be tested against the orthogonal quarter-planes defined by the other two axes of other timestamps. An example is given in Figure 4. The V-ray of one timestamp (orange) is tested against the quarter-plane defined by the T- and E-rays (T-E-quarter-plane) of a second timestamp (yellow). To allow for retroactive updates relative to the V-axis to "slide in between" existing values as in the two-time-dimensional case, the V-ray has to stop at the T-E-quarter-plane.

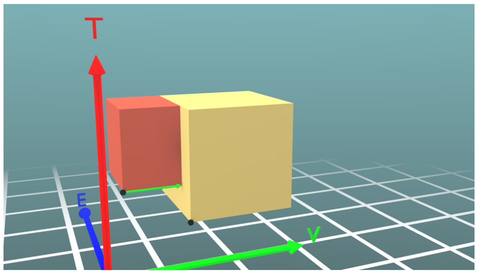


Figure 4. Expansion in V-direction of one timestamp's scope (orange) stops at T-E-quarter-plane of other timestamp's scope (yellow).

In general, the rule has to state whether a T-, V-, or E-ray has to stop at the respective orthogonal V-E-, T-E-, or T-V-quarter-planes. Like in the 2D case, a desired property is that the partitioning shall be independent of the versions' insertion order, i.e., of their timestamps' T component. Corresponding to the partitioning of the vertical line at t_j^T in Figure 1 (right) and Figure 2, this is the partitioning of a "high enough" intersecting V-E-plane, i.e., for a t_j^T greater than any occurring t_i^T. An example is given in Figure 5. Compared to Figure 4, the T component t_i^T of the i^{th} timestamp (orange) is less than that of t_{i+1}^T (yellow). In each case, the respective V and E components are identical. In both scenarios, however, the

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-383-4

final (i.e., with regard to T) partitioning along an intersecting V-E-plane (depicted in purple in Figure 5) is equal. In the first scenario ($t_i^T > t_{i+1}^T$, Figure 4), a V-ray *must* stop at a T-E-quarter-plane. In the second ($t_i^T < t_{i+1}^T$), an E-ray *must not* stop at a T-V-quarter-plane. Otherwise, the partitioning would not be consistent.

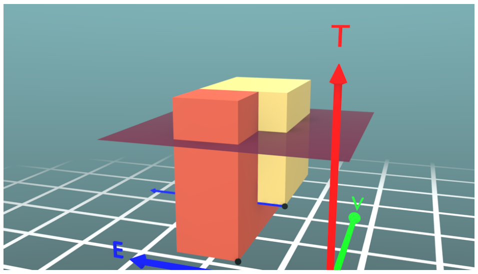


Figure 5. Expansion of E-ray must not stop at T-V-quarter-plane to yield the same partitioning on a V-E-plane (purple) as in Figure 4.

Figures 4 and 5 depict the case where $t_i^V < t_{i+1}^V$ and $t_i^E > t_{i+1}^E$. For reasons of symmetry, the same arguments apply to the case where $t_i^V > t_{i+1}^V$ and $t_i^E < t_{i+1}^E$. In fact, only the indexes (i and $i+1$) and thus the colors of the two scopes (orange and yellow) need to be swapped.

The other distinct case is $t_i^V < t_{i+1}^V$ and $t_i^E < t_{i+1}^E$. For $t_i^T < t_{i+1}^T$ this is already depicted in Figure 3. Here, all coordinates of timestamp i (orange) are less than those of timestamp $i+1$ (yellow). To achieve the same partitioning on a V-E-plane for $t_i^T > t_{i+1}^T$, T-rays *must not* stop at V-E-quarter-planes. Figure 6 shows an example. The T-ray emitted from the $(i+1)^{th}$ timestamp (yellow) does *not* stop at the V-E-quarter-plane defined by the i^{th} timestamp (orange). The dashed ovals mark the yellow scopes expansion reducing the orange scope. Again, the same argument applies to $t_i^V > t_{i+1}^V$ and $t_i^E > t_{i+1}^E$ and can be depicted by switching colors (orange and yellow) in Figures 5–6.

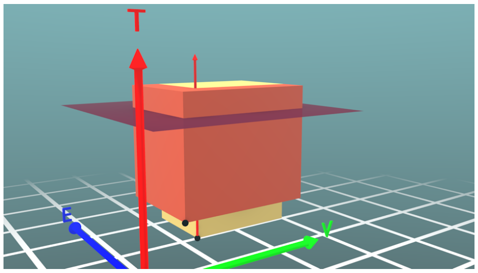


Figure 6. Expansion of a T-ray must not stop at V-E-quarter-planes to yield the same partitioning on a V-E-plane (purple) as in Figure 3.

Altogether, the Golden Rule for the three-time-dimensional case states that

- T-rays *must not* stop at V-E-quarter-planes,
- V-rays *must* stop at T-E-quarter-planes, and
- E-rays *must not* stop at T-V-quarter-planes.

The constraints for the choice of these rules are

- partitioning on a ”high enough” V-E-plane shall be independent from insertion order, i.e., transaction time T and
- retroactive updates relative to the V-axis shall ”slide in between” existing values.

IV. Implementation in Geodatabase

Our goal was to implement the time-related behavior of a geospatial reference system independently of any manufacturer. This approach ensures the portability between different database systems like Oracle or PostgreSQL, especially when managing time-related geospatial data. For that purpose, during the development of the presented nD temporal data management system, a transformational layer was introduced as an agent between application development and physical data storage of geospatial data (Figure 7). This layer provides a transformation between the temporal properties of the data schema (independent of explicit or implicit modeling of time) and its implementation, e.g., in terms of tables in an object-relational database.

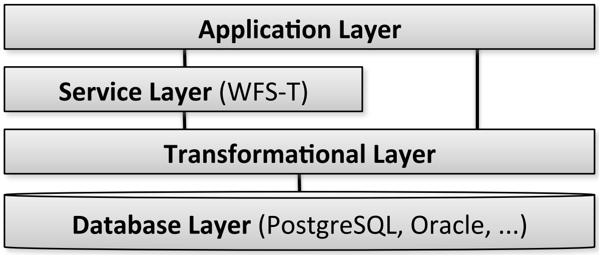


Figure 7. Layered architecture of the proposed approach.

On the application layer, GML is used to describe and request geospatial time-related data within the proposed data management system. A GML-encoded dataset combined with a request (select, insert, update, delete) is transformed into Structured Query Language (SQL) statements by the transformational layer.

The transformational layer is configured using an application schema provided as an Extensible Markup Language (XML) schema file (XML Schema Definition (XSD)) using the specifications of ISO 19109. This schema defines the rules for the automatic generation of SQL statements for the nD data storage.

For OGC-compliant data retrieval and manipulation, a Web Feature Service (WFS 2.0) was implemented as an interface to the transformational layer. It accepts geospatial data (in terms of GML 3.2.1) with multiple timestamps that is passed to the nD data management system where a consistent handling of the space-time-reference is ensured.

V. Integration into Simulation System

As presented in [2], we supply our simulation system clients with a shared simulation model. While the model is managed by a central (geo)database, each client uses an in-memory runtime simulation database to locally cache the model for real-time access (Figure 8). This combination of databases can be seen as a distributed database system that uses change notifications for replication and synchronization [12].

The simulation database itself is a snapshot database (compare Subsection II-B). Thus, replication from the central geodatabase must be performed with reference to a certain timestamp tuple provided by the user or an application component. To make the snapshot consistent, the entirety of replicate copies must always represent the same timestamp tuple. Thus, on changing the reference time, replication has to be updated. A naive approach is to unload all objects and then reload them with the new reference date. A more efficient approach is to only reload data that changed in between the previous and the new reference time. For this purpose, however, the central geodatabase must be able to (quickly) provide a list of all changes between two reference time tuples, e.g., by maintaining a queryable global change log.

Figure 8 shows an example scenario for the bitemporal case. The central geodatabase contains a simple forest model comprising a Tree class with a height attribute. A snapshot of the object a:Tree is replicated to the client using reference time (t^T=cur, t^V=2012). By using transaction time t^T=cur (the default access mode for most users), the last written value for valid time t^V=2012 is replicated. Thus, the height of a:Tree's snapshot within the simulation database is $16.0m$.

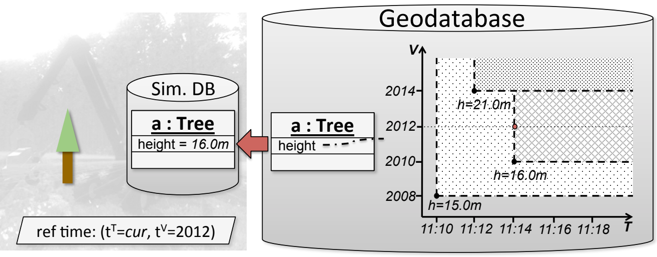


Figure 8. Bitemporal example: Replicating a snapshot of a tree object from a central geodatabase with reference time (t^T=cur, t^V=2012).

Besides replicating data from the central geodatabase, changes within a simulation client can also be synchronized back. For that purpose, first, the local snapshot must be refreshed to represent the desired reference time for the change. Subsequently, the change is applied locally and an update transaction with the exact same time tuple is issued to the central geodatabase. For the same reason, the simulation database must be considered read-only when the snapshot represents historic versions regarding transaction time (i.e., $t^T \neq$cur). Here, changes cannot not be resynchronized as transaction time does not support retroactive updates.

An example for updating the previously replicated a:Tree object is given in Figure 9.

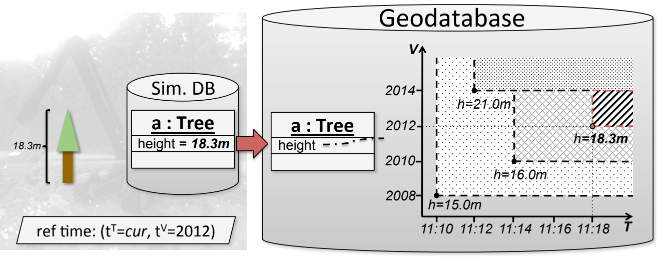


Figure 9. Bitemporal example: Updating a tree object's height within the central geodatabase for reference time (t^T=cur, t^V=2012) (transaction executed at 11:18h).

The tree's height actually measured $18.3m$ in 2012 so it must be updated. While keeping the same reference time (t^T=cur, t^V=2012), the new value is resynchronized to the geodatabase. It is inserted with the chosen valid time t^V=2012 and the system's current time for transaction time – in this case t^T=11:18.

VI. Distributed Parallel Data Management

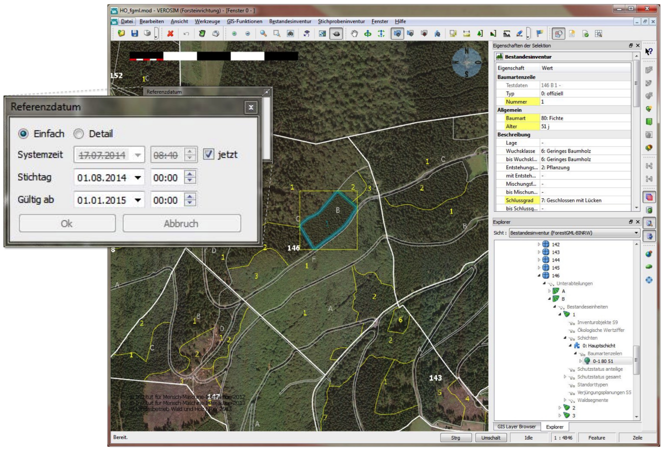
Based on the presented nD-versioning approach, a concept for distributed parallel data management was developed. In this scenario, the master copy of the considered geodata is managed in a central temporal database while replicate copies are distributed to several temporal working databases, on-demand. The latter are distributed over remote sites, e.g., at different contractors (or departments) of the main organization, with limited or no direct connectivity to the central site. A WFS-T (Figure 7) at the master site is used for data exchange. Note, however, that these working databases are *not* the simulation databases from Section V but rather additional temporal geodatabases.

For data acquisition (check-out), depending on connectivity, a contractor either directly queries the WFS-T or delegates his request to the organization. Given authorization, the data is marshaled to a GML representation comprising all its versions. In the GML representation, objects and properties are associated with timestamps in terms of XML attributes. To allow for XML validation, an extended schema specifying these attributes is issued by the WFS-T. On remote site, the GML data is unmarshaled (i.e., imported) into the working database. Altogether, this allows remote sites to access the full history of the acquired data – not only a snapshot. The contractor can perform his tasks on the data within his working database. To transmit his results back to the central site (check-in), changes are tracked within the working database. They are marshaled to a similar GML representation, which, however, only contains the changed objects, as well as the associated timestamps and their full history. This GML dataset is sent to the WFS-T – as before, either directly or indirectly – for check-in into the master database. Furthermore, the concept can also be applied in a cascade. Here, the contractor itself distributes the data once again to 2^{nd} tier working databases among his employees using a local WFS-T.

To allow different contractors to work in parallel on the same data, it can be replicated multiple times. For that purpose, it is marked with so-called *process control objects* within the central database. In contrast to locking, this allows for longterm processes to be performed in parallel without mutual exclusive access. Process control objects also define the reason for a check-out. Based on existing markers, further check-out requests can be granted for non-exclusive processes. This process-based approach allows for a flexible concurrency control within the distributed data management scenario. Due to this optimistic approach, however, change conflicts may arise when the same data item is changed by different contractors. Using the temporal information, this can be detected on check-in time by analyzing changes that occurred since the contractor's check-out by comparing the available full history of the changed object. Additionally, besides temporal data, metadata describing the job and the contractor is managed allowing to determine responsibilities for changes. For that purpose, before the actual check-in, a simulated check-in is performed: The corresponding check-out is repeated to a local helper database where the contractor's changes are applied. Here, conflicts can be detected and either be directly resolved or referred to the contractor for resolution. When no conflicts occur, the actual check-in is performed and the corresponding process control marker is removed.

VII. Applications

The research project Virtual Forest [13][14] is one of the primary applications for the presented approach. One of the core ideas of this project is a consistent, shared data model and data management in the Virtual Forest database. Provided to all stakeholders in this field, it facilitates the exploitation of know-how and synergies. Furthermore, it supports the transfer of industrial automation techniques to the forest industry. Note, however, that the presented techniques are not restricted to this context. Figure 10 shows one of the realized applications using the presented approach. The user interface to set the reference times ("Referenzdatum") is shown in detail. It allows to set transaction ("Systemzeit"), valid ("Stichtag") and effectivity time ("Gültig ab"). Transaction time can be set to *current* ("jetzt").



HO_fgml.mod - VEROSIM (Forsteinrichtung) - [Fenster 0 -]

Referenzdatum

● Einfach ○ Detail

Systemzeit ~~17.07.2014~~ ~~08:40~~ ☑ jetzt

Stichtag 01.08.2014 00:00

Gültig ab 01.01.2015 00:00

Ok Abbruch

Eigenschaften der Selektion

Bestandesinventur

Eigenschaft	Wert
Baumartenzeile	
Testdaten	146 B 1 -
Typ	0: offiziell
Nummer	1
Allgemein	
Baumart	80: Fichte
Alter	51 j
Beschreibung	
Lage	-
Wuchsklasse	6: Geringes Baumholz
bis Wuchskl...	6: Geringes Baumholz
Entstehungs...	2: Pflanzung
mit Entsteh...	-
Mischungsf...	-
bis Mischun...	-
Schlussgrad	7: Geschlossen mit Lücken
bis Schlussg...	-

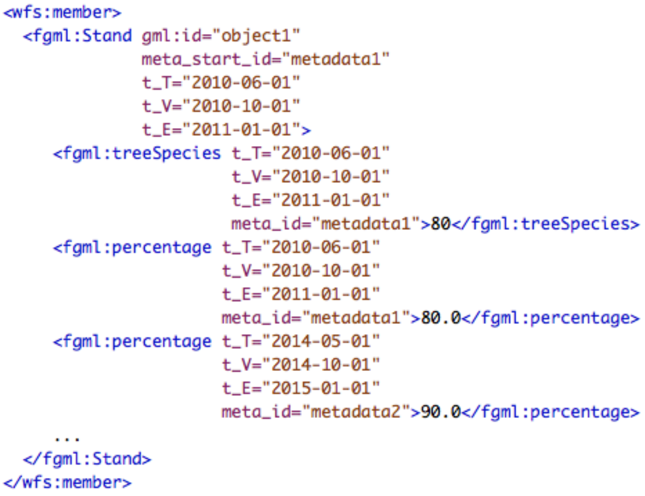
Explorer

Sicht : Bestandesinventur (ForestGML-BINRW)

Figure 10. A forest inventory tool based on the presented approach - the reference date for the local snapshot can be specified by the user.

For the Virtual Forest database, currently, the three aforementioned time dimensions are utilized. Transaction time is used to model the technical change history. It allows to determine when forestal data was changed, thus, documenting work processes. The second time dimension is valid time. In particular for individual tree objects, it is used to directly model the actual age. Id est, a tree's "birth" is modeled as a new tree object inserted with the corresponding valid time. Any change to its properties is equally updated, e.g., a height measured at a certain point in time (see example in Section V). In stand inventory, forests are not modeled as individual trees but in terms of so-called *stand units*. A stand unit object combines a surface geometry with aggregated properties like 75% spruce with average age of $80y$ and average height of $30m$. All associated updates to this stand inventory data are typically bound to a certain valuation date within the considered year, e.g., October 1st. Thus, all updates use this date as valid time, although the actual values are typically recorded over a longer period of time around this date. Effectivity time is used to model the point in time when new values shall become the official representation of the forest. Updating huge forest datasets using fieldwork or automated processes (as developed in the Virtual Forest project) can take a longer period around the valuation date. By using an effectivity time well after this period (e.g., next New Year's Day), no intermediate results are visible to other users evaluating the currently effective data.

In the context of the Virtual Forest project, a GML application schema called ForestGML was developed to consistently model forest-related data. To add the necessary multiple timestamps on a per-attribute basis, standard GML or WFS concepts like `<wfs:FeatureCollection timeStamp="...">` do not suffice. Thus, to embed nD temporal reference information and corresponding metadata into this (or any other) application schema, it is extended by XML attributes. Figure 11 gives a simplified example of a stand unit object including all its historical attribute versions for a check-out process. The object itself is extended by the creation timestamps (t^T=2010-06-01, t^V=2010-10-01 and t^E=2011-01-01). The object's creation is described in the metadata object referenced by `meta_start_id`. The tree species code and area percentage attribute both were set during object creation yielding the same reference time and metadata object. However, the percentage value was changed at a later point in time (t^T=2014-06-01, t^V=2014-10-01 and t^E=2015-01-01) in the context of another job described by `metadata2`.



```
<wfs:member>
  <fgml:Stand gml:id="object1"
              meta_start_id="metadata1"
              t_T="2010-06-01"
              t_V="2010-10-01"
              t_E="2011-01-01">
    <fgml:treeSpecies t_T="2010-06-01"
                      t_V="2010-10-01"
                      t_E="2011-01-01"
                      meta_id="metadata1">80</fgml:treeSpecies>
    <fgml:percentage t_T="2010-06-01"
                     t_V="2010-10-01"
                     t_E="2011-01-01"
                     meta_id="metadata1">80.0</fgml:percentage>
    <fgml:percentage t_T="2014-05-01"
                     t_V="2014-10-01"
                     t_E="2015-01-01"
                     meta_id="metadata2">90.0</fgml:percentage>
    ...
  </fgml:Stand>
</wfs:member>
```

Figure 11. Simplified example using the ForestGML application schema extended by nD temporal information and metadata.

Finally, using the presented approach for distributed parallel data management, inventory jobs can be distributed among several contractors (or employees of one contractor). Processes modeled using process control objects comprise inventory jobs, geometry revisions, or annual automatic growth extrapolations. Occurring change conflicts, e.g., due to inaccurately performed jobs, can be resolved as described in Section VI.

VIII. Conclusion and Future Work

All in all, we presented a novel approach to manage temporal geodata with multiple time dimensions. In particular, it exceeds the bitemporal scenario adding further flexibility in temporal data modeling. Furthermore, we developed a new approach to use point representations for several time dimensions by defining a "Golden Rule" to resolve ambiguities. Based on this nD-versioning, a flexible approach for distributed parallel data management is presented. In the end, prototypical implementations and the practical usage for forestry applications already prove the feasibility of the approach.

As presented in [12], our approach for distributed database synchronization with multiple clients relies on change notifications. This concept has to be extended for time dimensions allowing retroactive or proactive updates. On the part of the central (geo)database, notifications would need to comprise the timestamp tuple corresponding to the notified change. The component for distributed synchronization, in turn, would need to consider these timestamps. The particular notification handling strategy requires further research. Likewise, for the multiple time dimension case, the efficient updating of local replicate copies within a simulation database when changing the considered reference time has to be analyzed.

ACKNOWLEDGMENT

Virtual Forest: This project is co-financed by the European Union and the federal state of North Rhine-Westphalia, European Regional Development Fund (ERDF). Europe - Investing in our future.

REFERENCES

[1] "The Consensus Glossary of Temporal Database Concepts," URL: http://people.cs.aau.dk/~csj/Glossary/index.html [accessed: 2015-01-03].

[2] M. Hoppen, M. Schluse, J. Rossmann, and B. Weitzig, "Database-Driven Distributed 3D Simulation," in Proceedings of the 2012 Winter Simulation Conference, 2012, pp. 1–12.

[3] J. Rossmann, A. Bücken, and M. Hoppen, "Semantic World Modelling and Data Management in a 4D Forest Simulation and Information System," ISPRS 8th 3DGeoInfo Conference & WG II/2 Workshop, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-2/W2, 2013, pp. 65–72.

[4] R. Elmasri and S. B. Navathe, Database Systems: Models, Languages, Design, And Application Programming, 6th ed. Prentice Hall International, 2010.

[5] T. Myrach, Temporale Datenbanken in betrieblichen Informationssystemen, 1st ed. Wiesbaden: Teubner Verlag, 2005.

[6] J. Clifford and A. U. Tansel, "On an algebra for historical relational databases: two views," ACM SIGMOD Record, vol. 14, no. 4, May 1985, pp. 247–265.

[7] N. Pelekis, B. Theodoulidis, I. Kopanakis, and Y. Theodoridis, "Literature Review of Spatio-temporal Database Models," The Knowledge Engineering Review, vol. 19, no. 3, Sep. 2004, pp. 235–274.

[8] P. van Oosterom and J. Stoter, "5D data modelling: full integration of 2D/3D space, time and scale dimensions," in Proceedings of the 6th international conference on Geographic information science. Zurich, Switzerland: Springer-Verlag, Sep. 2010, pp. 310–324.

[9] W. Siabato, C. Claramunt, M. A. Manso-Callejo, and M. A. Bernabé-Poveda, "Time Bliography: A Dynamic and Online Bibliography on Temporal GIS," Transactions in GIS, vol. 18, no. 6, Dec. 2014, pp. 799–816.

[10] ISO/TC 211 Geographic information/Geomatics, "ISO 19100," URL: http://www.isotc211.org [accessed: 2015-01-03].

[11] "Open Geospatial Consortium (OGC)," URL: http://www.opengeospatial.org [accessed: 2015-01-03].

[12] M. Hoppen and J. Rossmann, "A Database Synchronization Approach for 3D Simulation Systems," in DBKDA 2014,The 6th International Conference on Advances in Databases, Knowledge, and Data Applications, A. Schmidt, K. Nitta, and J. S. Iztok Savnik, Eds., Chamonix, France, 2014, pp. 84–91.

[13] J. Rossmann, M. Schluse, and A. Bücken, "The virtual forest - Space- and Robotics technology for the efficient and environmentally compatible growth-planing and mobilization of wood resources," FORMEC 08 - 41. Internationales Symposium, 2008, pp. 3 – 12.

[14] J. Rossmann, M. Schluse, R. Waspe, and R. Moshammer, "Simulation in the Woods: From Remote Sensing based Data Acquisition and Processing to Various Simulation Applications," in Proceedings of the 2011 Winter Simulation Conference, S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, Eds., 2011, pp. 984 – 996.

Personalized Trajectory Reconstruction Problem with Low-Sampling Data

Edison Ospina, Francisco Moreno
Departamento de Ciencias de la Computación y la Decisión
Universidad Nacional de Colombia. Facultad de Minas.
Medellin, Colombia
e-mail: ecospinaa@unal.edu.co; fjmoreno@unal.edu.co

***Abstract*—A huge amount of trajectory data can be derived from GPS equipped devices and location based services. However, the trajectory data are low-sampled (i.e., have a low and irregular sampling rate). In this paper, the problem of Personalized Route/Trajectory Reconstruction is reviewed when low-sampling data are considered or user criteria are incorporated in the reconstruction. Research work in route planning systems and the most used routing algorithms are also analyzed in order to suggest research directions that include the level of personalization or uncertainty management as a way to predict/complete a low-sampling trajectory.**

Keywords-Personalized Route Planning Systems; Location-Based Services; Trajectory Reconstruction; Uncertainty.

I. INTRODUCTION

Being able to choose the most convenient route to travel from one place to another is a desirable possibility when planning activities. For example, tourists usually ask for the best routes for visiting attractive places. Fields such as logistic, traffic control, and advertising also demand solutions in this regard in order to meet a variety of requirements, such as quality of road, cost of fuel, route availability, and user preferences, among others [1][2][3]. Several authors have recently been focused on the incorporation of user preferences and multi-criteria decision making aspects in light of the route personalization [3]. Other approaches have used GPS data representing historical movements of users based on individual [5] or collective behavior [6]. The resulting routes are usually closer to the typical ones actually followed by users than those suggested by route planners as optimal (the shortest and fastest) [7][8].

According to the literature reviewed, the terms *Route Finding Problem* and *Path Finding Problem* are used interchangeably. Other term related to the Route Finding Problem (RFP) is *Routing* (or Route) *Planning Systems* (RPS). The request for a route to travel from one place to another in the RFP is considered the pair for finding a trajectory between low-sampled points. Therefore, the reviewed research works are analyzed in relation to the RFP, paying special attention to those taking into account *user criteria* or *low-sampling-rate data* (i.e., when the time interval between consecutive GPS points of some trajectories is higher than a given threshold) [1]. When low-sampling-rate data is present, the reconstruction of trajectories may be needed, i.e., the description of the movement of the object between the two points where no data points are available to know where the object is while travelling.

The rest of this paper is organized as follows: Section II describes routing planning systems; Section III describes personalization, i.e., incorporation of user preference criteria as a way to deal with the trajectory reconstruction problem; Section IV addresses the reconstruction of trajectories under low-sampling-rate data, and Section V concludes the paper and proposes future work.

II. ROUTING PLANNING SYSTEMS

RPS are commonly recognized as decision support systems [9][10]. These systems sometimes are referred to as geo-related decision support tools [10]. In Table I, some variations of the term referring to RPS are presented. Conventional solutions to RFP are limited because the routing is based on just one dimension (criterion): the cost [11][12][13]. Many definitions include, explicitly or implicitly, the notion of personalization, suggesting that user interaction is required. Recent researches have been carried out to improve these models through their personalization and the incorporation of multi-criteria decision making including preference models [3][4].

TABLE I. COMMON TERMS REFERRING TO ROUTING PLANNING SYSTEMS.

Author	Term	Comment
[4]	Routing systems	Routing systems aim to help users on finding the optimal path to their destination regarding travel distance, travel time, among other criteria.
[10]	Personalized user-centric route finding	A personalized user-centric route finding application incorporates user preferences and the environmental features around a user. User preferences and environmental features are the key elements to assess a route.
[3]	Personalized route planning systems	A personalized route planning system provides a route based on minimizing a combination of user defined criteria such as travel distance, travel time, the number of traffic lights, and road types.
[11]	Route guidance systems	Route guidance systems refer to all the factors considered before and during a trip to choose or adjust a route. Route guidance systems are recognized as a fundamental component of intelligent transportation systems.

ISBN: 978-1-61208-383-4

A brief schema review of the RFP in RPS is shown in Figure 1. The RPS are supported by *Routing Planning Algorithms*. When the personalization is included, incorporating preferences or decision strategies originates the concept of *Personalized Routing Planning Systems*.

The classical algorithm for RFP based on the shortest path issue was proposed by Dijkstra [14] and it has been used widely to find the shortest path between an origin vertex and a destination vertex in a weighted graph, exploring the entire graph to determine the lowest cost route. Similarly, the A* algorithm (a modification of Dijkstra's algorithm) finds the optimal path using an appropriate heuristic that defines which is the best node to be visited next (it avoids explore the entire graph) based in the lowest heuristic cost [15], e.g., some of the Minkowski metrics [16]. All of these early approaches are based on algorithms that use an *edge cost*, i.e., they performed a one-dimensional analysis. For this reason, these algorithms are inadequate or incomplete since users generally have different purposes and they do not share the same movement behavior, highlighting the need to *personalize* and allow the user to interact with RPS.

III. Personalization

The technology-based definition of personalization provided by the Personalization Consortium (2005) is "the use of the technology and customer information to tailor electronic commerce interactions between a business and each individual customer". An experiment conducted by Golledge [17] showed that the criteria used by humans to deal with path selection problems may be a complex task that covers a wide spectrum of choices. The routes were determined using criteria selection such as shortest distance and fewest turns. Variables such as orientation and the possibility of retracing the route (i.e., interchange the origin and the destination) were also studied to determine the change of the user route criteria selection when traveling in one direction or the other. To illustrate the above problem, two possible routes between an origin O and a destination D are shown in Figure 2. The route O-C-D is usually suggested by common RPS without considering the probability of a traffic jam or local restrictions for moving between streets.

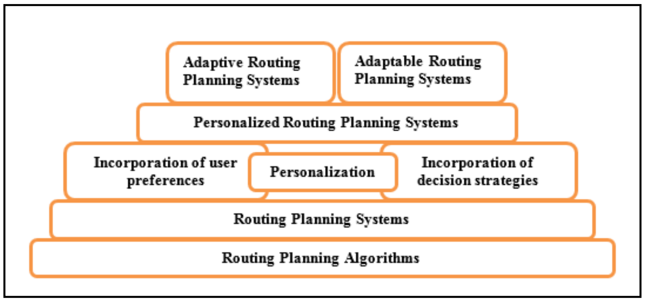

Adaptive Routing Planning Systems
Adaptable Routing Planning Systems
Personalized Routing Planning Systems
Incorporation of user preferences
Personalization
Incorporation of decision strategies
Routing Planning Systems
Routing Planning Algorithms

Figure 1. Schema review of the RFP according to personalization in RPS.

However, most users would select the route O-A-B-D even though path O-C-D has the minimum distance, because more points of interest (POI) can be found along it (supermarkets, parks, or gasoline stations). This is evidenced by Duckham and Kulik [18], showing how a *simple path* solution offers considerable advantages over shortest paths in terms of ease of description and execution. Several researchers have stated the importance of the personalization when solving routing planning tasks [3][5][9].

The goal of personalization is the automatic adaptation of an information service in response to the *implicit* or *explicit* needs of a specific user [9]. That is, the automatic identification of preferences from the user movement behavior history [7][8] or explicit requests of the user [3][10]. Also, Fischer [19] stated that personalization can be described by *adaptable* and *adaptive* methods, and Oppermann [20] gives the following definition to those terms: in adaptable systems the user controls the adaptation process whereas in adaptive systems the process is automatic, i.e., without user intervention. Nadi and Delavar [3] define adaptable and adaptive personalized route guidance systems in the context of RPS. Examples of adaptable [3][10] and adaptive [21][22] RPS can be widely found in the literature.

In [13], static and dynamic systems, deterministic and stochastic systems, reactive and predictive systems, and centralized and decentralized systems are distinguished. In [11], descriptive and prescriptive guidance and static and dynamic guidance are reviewed. In [12], route guidance systems are classified as infrastructure-based and infrastructure-less systems. Infrastructure-based systems are based on two components: i) hardware devices deployed in streets/roads and ii) computer systems installed in moving objects (e.g., a GPS). Infrastructure-less systems require only the second component. Personalization can also be defined in terms of user route choice criteria.

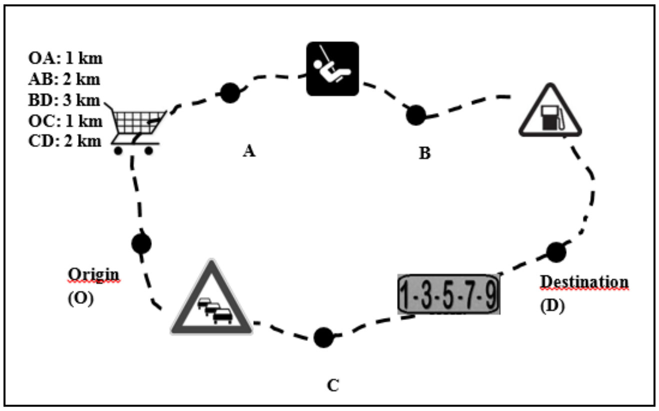


Figure 2. Problem of route finding in a road network.

A special issue of the personalization in RPS is the characterization and incorporation of several criteria, e.g., route length or travel time. Table II shows some of them classified as quantitative (they are measured from a map or any other source) and qualitative (they are no-numeric criteria that are ranked according to the impact on the user). Previous research [3][4][23] found that route selection criteria can be grouped into four general criteria: speed (time, distance), safeness, simplicity, and attractiveness (POIs-based scenic path):

A. Speed: Distance

Distance is normally considered the most important criterion for route choosing. Even without route planning systems, the path with the shortest distance is intuitively chosen with a minimum previous knowledge of the RN structure (however, the presence of known POIs may lengthen the road trip. See *attractiveness*).

B. Speed: Time

Time is a variable that depends of several factors such as length (time is directly proportional to the length of road), average speed (higher in main avenues than in small streets), quality of roads, and weather conditions (e.g. when it rains, travel time is higher due to traffic conditions derived from it) or quality of traffic as described in [4].

C. Safeness

It groups a series of criteria based on characteristics (bike lane availability, area safeness, night lighting, traffic level), possibilities (lack of busy intersections, public transport, and roundabouts), and features of the road (presence or lack of pavement, slope angle) [23].

D. Simplicity

The simplest path is based on the idea that the turns imply reductions of velocity and unnecessary maneuvers. Thus, the path is “better” if it has fewer turns [4]. Moreover, the description of the path is easier when a simplest path approach is followed, as the explanation, depiction, understanding, memorizing, and/or execution of it [18], which is useful for users who are navigating through an unfamiliar geographic environment.

TABLE II. QUANTITATIVE AND QUALITATIVE CRITERIA OF RFP.

Author	Criteria	Quantitative	Qualitative
[3][4]	Distance, Travel Time	x	
[4][8][23]	Traffic	x	
[3][4][18]	Costs of Turns/ Simplest Paths	x	
[24][25]	Number of Scenic Landscapes / POIs	x	
[3]	Number of Junctions, Travel Reliability, Directness, Road Width, Number of Stop Signs	x	
[3]	Quality of Road, Type of Road		x

E. Attractiveness

Variables such as distance, time, or turns are common route criteria for navigating a street network, but computation of the most scenic route is a special issue [26]. The scenic-path notion is defined from the touristic perspective. The main idea is to travel from A to B trying to visit as much touristic places as possible and minimizing route length at the same time. The cost is the number of touristic attractions between the two points (for instance, the streets with a considerable number of POIs have the lowest cost). A modification of the cost is required before the execution of a shortest path algorithm if the goal is to find a route that traverses as much POIs as possible and, at the same time, the shortest route between two POIs.

Figure 3 exhibits a section of Guarne, a small town in Colombia, with a route between two points using the shortest path algorithm. Figure 3(a) shows the minimum distance between point A and B. Figure 3(b) shows the route with the minimum travel time between point A and point B. Figure 3(c) shows the route between the two points using the simplest path approach. The turns in the path are less in the latter, even though the whole path may be longer. Figure 3(d) shows the route using the scenic path approach: the route is draw along the street nearest to the town river where touristic attractions are present (restaurants, beach games, etc.).

Figure 3 shows how a path may vary when different criteria are considered. Users not always choose the shortest route. This set of exercises provides evidence that route selection is a process that requires support of decision strategies and preference models to back personalization.

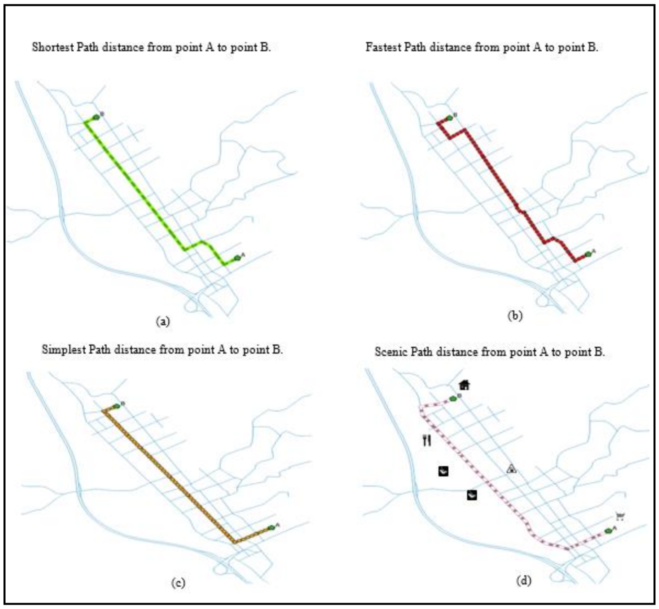

Shortest Path distance from point A to point B.
Fastest Path distance from point A to point B.
(a)
(b)
Simplest Path distance from point A to point B.
Scenic Path distance from point A to point B.
(c)
(d)

Figure 3. Different route finding criteria from point A to point B.

IV. PERSONALIZED ROUTE FINDING BASED ON TRAJECTORIES

The RFP reviewed here is the reconstruction of low-sampling trajectories. To solve this problem, pattern-based and greedy searches approaches has been considered (Preference-based Greedy search, NaïVe Greedy search, Pattern+Greedy search) [27]. Pattern-based approaches allow *offline* processing of historical trajectory data to discover mining patterns and infer routing information [1], while greedy search approaches make optimal local choices at every decision stage, providing a dynamic/*online* recommendation on the best immediate location to be visited [27]. Most of these procedures deal with a general mining/prediction problem over historical trajectories [6][8][27][28]. In the reviewed works, the personalization is based on the trajectory history data of a particular user.

A. *Route Planning based on GPS trajectories*

In [29], the problem of searching the *k-Best Connected Trajectories* (*k*-BCT) is addressed. A small set of locations (queried points) is given as an input to an incremental k-NN (K-Nearest Neighbor) based algorithm, which progressively retrieves trajectories nearest to each location, using best-first and depth-first k-NN algorithms. The quality of the connection between locations provided by the discovered trajectories is given by a similarity measure. A dataset of Beijing collected by the Microsoft GeoLife Project was used to analyze the efficiency of the KNN algorithm, showing a better search performance if the best-first k-NN algorithm is chosen. In [6], the problem of discovering the *most popular* route between two given locations using historical user trajectories is addressed. A *Coherence Expanding Algorithm* is proposed for mining users movements together with a popularity indicator. Then, an algorithm for searching the most popular route given two locations is applied. Considering 276 truck trajectories used in Athens and applying the proposed algorithm, the most popular routes were identified. Then, these findings were compared against those obtained with the shortest path approach. In [5], a *Pattern-aware Personalized routing framework* (PPT) is proposed using a two-step method to compute personalized routes. First, a set of frequent road segments is derived from a user historical trajectories database to construct a familiar RN followed by a specific user. Then, while a route is computed between a specific source and a destination, a second algorithm is proposed to discover the top-k personalized routes connecting some of the segments that a user has previously traveled. The algorithms were tested using a real trajectory dataset from one user in Kaohsiung, Taiwan. The algorithms derived the top-k personalized routes that approximate the real top-k personalized routes. In [8], smart driving directions are mined from taxi drivers' experience. A routing algorithm is proposed to provide the fastest route from a given origin to a given destination. Thus, a time-dependent graph is built where nodes are recognized as landmarks, i.e., road segments traversed by a significant number of taxis and edges represent taxi routes between landmark roads. This demonstrates that about 16% of time can be saved with this method compared to speed-constraint and real time traffic-based methods. In [7], fast routes are mined from taxi traces and are customized for a particular driver behavior. A mobile device learns about the user driving behavior thanks to the user driving routes and finds the fastest route. This model outperforms the previous work [8]. In [27], the construction of a preferred route using location check-in data are done based in the popularity of a certain route and the preferences ranked by a set of users. The goal is to build a trajectory where the reconstruction meets the preferred locations to be visited by a *group of persons* using Gowalla check-in data and a Pattern+Greedy method (this combination of Pattern and Greedy route search outperforms both methods when used separately). Similarly, in [28], the top-k Trajectories are extracted from interesting regions with higher scores (attractiveness) mined from historical GPS trajectories. A framework for trajectory search called Pattern-Aware Trajectory Search (PATS) is developed, which includes an off-line pattern discovery module and an online pattern-aware trajectory search module. This framework only searches for the top-k maximal trajectories with higher scores according to the number of interesting regions and does not infer new routes.

B. *Uncertainty in Trajectories*

When a trajectory is reconstructed, its uncertainty should be considered. Uncertainty from different sources is evidenced by Kuijpers and Moelans [30]: i) Accuracy of the GPS observation and ii) the uncertainty derived from low sampled points of a trajectory. Those are also referred as measurement and sampling errors [31]. Previous works [5][6][8][29] relied on high-sampled trajectories; however, the effectiveness of inferred routes is poor due to its inadequate management of low-sampling trajectories where uncertainty is reflected. The causes for *low-sampling* trajectories include the lack of users sharing their position or taking geo-tagged photos from every place and every second. This is due to the privacy concerns publishing personal location data to potentially untrustworthy service providers may pose [32]. Research works has been carried out to preserve publishing data of a moving object to a third party for data analysis purposes [33][34]. Privacy-preserving techniques has been studied based on false location [35], space transformations [36], or spatial cloaking [37]. However, those works are not aimed to reduce low-sampling directly. Instead, they provide privacy-preserving techniques to promote location, sharing information.

The main features of the trajectories regarding to uncertainty are highlighted in [38]:

1) *Spatial Biases:* The locations of data points in two trajectories are different, i.e., two similar trajectories can be depicted by means of different location data points.

2) *Temporal Biases:* The occurrence time of two trajectories are different, i.e., two similar trajectories visiting the same POIs could be done in two different time periods.

3) *Silent Durations:* The time periods when no data points are available to describe the movements of the users.

Relevant data are missing during silent durations. User movement criteria can fulfill partially those silent durations. For the best of our knowledge, the low-sampling-rate trajectory reconstruction problem has not considered the user preferences. We strongly believe this is a rich research area with application in several domains. For example, for location-based advertising, it might mean the possibility of advertising strategies based on data about routes followed by users from a POI A to a POI B.

Several studies [27][39][40] infer routes from a sequence of POIs but a detailed route between two consecutive POIs is not specified. The underlying assumptions of these works are that the user movement is free. However, the infrastructure, e.g., buildings, may be considered to obtain a reduced overall uncertainty and inaccuracy in the data. In [2], a Route Inference framework based on Collective Knowledge (RICK) is developed. Given a set of locations and a time span, a two-step method is followed: first, a “routable graph” is built and, then, the top-k routes according to the route inference algorithm are constructed. Two real dataset are used: registers of Foursquare check-in application used in Manhattan and trajectories used in Beijing. The aim is to demonstrate the effectiveness and efficiency of RICK. In [1], the problem of reducing uncertainty for a given low-sampling-rate trajectory is addressed. Historical data are used to discover popular routes as an estimation of low-sampling trajectories. A real trajectory dataset generated by taxis in Beijing in a period of three months is used to validate the effectiveness of their proposal and shows higher accuracy than the existing map matching [41].

V. Conclusions and Future Work

The trajectory reconstruction problem is still an open research issue, especially what is related to uncertainty due to low-sampling data and incorporation of user preferences. Simple linear interpolation, as a method of reconstruction, does not represent users real movement because they move according to a certain criteria such as time or the amount of touristic/scenic places. Indeed, the reconstruction of trajectories using user preferences is expressed as a need in recent research works [38][42]. As far as we know, there are no works that involve several criteria as a way to reconstruct low-sampling trajectories. This approach can be enhanced by the restriction of the movement in a RN [43] and methods to predict the location of moving objects in a RN [44]. Moreover, the current availability of GPS loggers gathered from mobile devices are useful in a variety of ways to make driving better [45], but effective usage of the huge amount of data is still a challenge [46]. Considering the different possibilities of user criteria reconstruction of trajectory and the huge amount of low-sampling data, data analysis tasks related to these possibilities of reconstruction can be conducted. Therefore, analytic results over reconstructed trajectories can vary if different criteria of reconstruction are used. For example, if a trajectory is reconstructed based on the criterion of minimize turns, the main avenues might be interesting for analysis tasks because those are the longest without deviations, but if the amount of POIs are used as a criterion of reconstruction, then the avenues nearest to tourist attractions might be the interesting ones.

References

[1] K. Zheng, Y. Zheng, X. Xie, and X. Zhou, “Reducing uncertainty of low-sampling-rate trajectories,” In Data Engineering (ICDE), 2012 IEEE 28th International Conference on, IEEE, Apr. 2012, pp. 1144-1155, doi: 10.1109/ICDE.2012.42.

[2] L. Wei, Y. Zheng, and W. Peng, “Constructing popular routes from uncertain trajectories,” In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Aug. 2012, pp. 195-203, doi: 10.1145/2339530.2339562.

[3] S. Nadi and M. Delavar, “Multi-criteria, personalized route planning using quantifier-guided ordered weighted averaging operators,” International Journal of Applied Earth Observation and Geoinformation, No. 3, Jun. 2011, pp. 322-335, doi: 10.1016/j.jag.2011.01.003.

[4] E. Silva and C. de Baptista, “Personalized path finding in road networks,” In Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on, vol. 2, IEEE, Sep. 2008, pp. 586-591, doi: 10.1109/NCM.2008.211.

[5] K. Chang, L. Wei, M. Yeh, and W. Peng, “Discovering personalized routes from trajectories,” In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, ACM, Nov. 2011, pp. 33-40, doi: 10.1145/2063212.2063218.

[6] Z. Chen, H. Shen, and X. Zhou, “Discovering popular routes from trajectories,” In Data Engineering (ICDE), 2011 IEEE 27th International Conference on, IEEE, Apr. 2011, pp. 900-911, 2011, doi: 10.1109/ICDE.2011.5767890.

[7] J. Yuan, Y. Zheng, X. Xie, and G. Sun, “Driving with knowledge from the physical world,” In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Aug. 2011, pp. 316-324, doi: 10.1145/2020408.2020462.

[8] J. Yuan, Y. Zheng, C. Zhang, and W. Xie, “T-drive: driving directions based on taxi trajectories,” In Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems, ACM, Nov. 2010, pp. 99-108, doi: 10.1145/1869790.1869807.

[9] L. McGinty and B. Smyth, “Shared experiences in personalized route planning,” In FLAIRS Conference, May. 2002, pp. 111-115.

[10] A. S. Niaraki and K. Kim, “Ontology based personalized route planning system using a multi-criteria decision making approach,” Expert Systems with Applications 36, no. 2, Mar. 2009, pp. 2250-2259, doi: 10.1016/j.eswa.2007.12.053.

[11] W. Herbert and F. Mili, “Route guidance: state of the art vs. state of the practice,” In Intelligent Vehicles Symposium, 2008 IEEE, IEEE, Jun. 2008. pp. 1167-1174, doi: 10.1109/IVS.2008.4621289.

[12] M. Khanjary and S. Hashemi, “Route guidance systems: review and classification,” In Proceedings of the 6th Euro American Conference on Telematics and Information Systems, ACM, May. 2012, pp. 269-275, doi: 10.1145/2261605.2261646.

[13] E. Schmitt and H. Jula, “Vehicle route guidance systems: Classification and comparison,” In Intelligent Transportation Systems Conference, 2006. ITSC'06, IEEE, Sep. 2006, pp. 242-247, doi: 10.1109/ITSC.2006.1706749.

[14] E. Dijkstra, “A note on two problems in connexion with graphs,” Numerische Mathematik 1, no. 1, 1959 pp. 269-271, doi: 10.1007/BF01386390.

[15] P. Hart, N. Nilsson, and B. Raphael, “A formal basis for the heuristic determination of minimum cost paths,” Systems Science and Cybernetics, IEEE Transactions on 4, no. 2, Jul. 1968, pp. 100-107, doi: 10.1109/TSSC.1968.300136.

[16] J. Van de Geer, “Some aspects of Minkowski distances,” Department of Data Theory, Leiden University, 1995.

[17] R. Golledge, “Path selection and route preference in human navigation: A progress report,” Springer Berlin Heidelberg, pp. 207-222, 1995, ISBN: 978-3-540-60392-4, doi: 10.1007/3-540-60392-1_14.

[18] M. Duckham and L. Kulik, “‘Simplest’ Paths: Automated Route Selection for Navigation,” In Spatial Information Theory. Foundations of Geographic Information Science, Springer Berlin Heidelberg, 2003, pp. 169-185, doi: 10.1007/978-3-540-39923-0_12

[19] G. Fischer, “Shared knowledge in cooperative problem-solving systems - Integrating adaptive and adaptable components,” In Schneider-Hufschmidt, M., Kuehme, T., & Malinowski, U., (Eds.) Adaptive User Interfaces, Principles and Practice. Elsavier Science Publishers, Amsterdam, pp. 49 – 68, 1993.

[20] R. Oppermann, Adaptive user support: ergonomic design of manually and automatically adaptable software. CRC Press, 1994. ISBN: 0-8058-1655-0.

[21] G. Pang and K. Takabashi, “Adaptive route selection for dynamic route guidance system based on fuzzy-neural approaches,” Vehicular Technology, IEEE Transactions on 48, no. 6, Nov. 1999, pp. 2028-2041, doi: 10.1109/25.806795.

[22] S. Rogers, C. Fiechter, and P. Langley, “An adaptive interactive agent for route advice,” In Proceedings of the third annual conference on Autonomous Agents, ACM, Apr. 1999, pp. 198-205, doi: 10.1145/301136.301193.

[23] H. Hochmair, “Towards a classification of route selection criteria for route planning tools,” In Developments in Spatial Data Handling, Springer Berlin Heidelberg, 2005, pp. 481-492, doi: 10.1007/3-540-26772-7_37.

[24] M. Kenteris and D. Gavalas, “Near-optimal personalized daily itineraries for a mobile tourist guide,” In Computers and Communications (ISCC), 2010 IEEE Symposium on, IEEE, Jun. 2010, pp. 862-864, doi: 10.1109/ISCC.2010.5546761.

[25] D. Gavalas, “Personalized routes for mobile tourism,” In Wireless and Mobile Computing, Networking and Communications (WiMob), 2011 IEEE 7th International Conference on, . IEEE, Oct. 2011, pp. 295-300, doi: 10.1109/WiMOB.2011.6085385.

[26] H. Hochmair and G. Navratil, “Computation of scenic routes in street networks,” In Proceedings of the Geoinformatics Forum Salzburg, Wichmann Verlag, 2008, pp. 124-133.

[27] H. Hsieh and C. Li, “Constructing trip routes with user preference from location check-in data,” In Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, ACM, Sep. 2013, pp. 195-198, doi: 10.1145/2494091.2494155.

[28] L. Wei and W. Peng, “Pats: A framework of pattern-aware trajectory search,” In Mobile Data Management (MDM), 2010 Eleventh International Conference on, IEEE, May. 2010, pp. 372-377, doi: 10.1109/MDM.2010.93.

[29] Z. Chen, H. Shen, X. Zhou, Y. Zheng, and X. Xie, “Searching trajectories by locations: an efficiency study,” In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, ACM, Jun. 2010, pp. 255-266, doi: 10.1145/1807167.1807197.

[30] B. Kuijpers and B. Moelans, “Analyzing trajectories using uncertainty and background information,” In Advances in Spatial and Temporal Databases, Springer Berlin Heidelberg, 2009, pp. 135-152, ISBN: 978-3-642-02981-3, doi: 10.1007/978-3-642-02982-0_11.

[31] D. Pfoser and C. Jensen, “Capturing the uncertainty of moving-object representations,” In Advances in Spatial Databases, Springer Berlin Heidelberg, 1999, pp. 111-131, ISBN: 978-3-540-66247-1, doi: 10.1007/3-540-48482-5_9.

[32] C. Chow and M. Mokbel, “Privacy of spatial trajectories,” In Computing with spatial trajectories, Springer New York, 2011. pp. 109-141, ISBN: 978-1-4614-1628-9, doi: 10.1007/978-1-4614-1629-6.

[33] O. Abul, F. Bonchi, and M. Nanni, “Never walk alone: Uncertainty for anonymity in moving objects databases,” In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, IEEE, Apr. 2008, pp. 376-385, doi: 10.1109/ICDE.2008.4497446.

[34] M. Gruteser and D. Grunwald, “Anonymous usage of location-based services through spatial and temporal cloaking,” In Proceedings of the 1st international conference on Mobile systems, applications and services, ACM, May. 2003, pp. 31-42, doi: 10.1145/1066116.1189037.

[35] J. Hong and J. Landay, “An architecture for privacy-sensitive ubiquitous computing,” In Proceedings of the 2nd international conference on Mobile systems, applications, and services, ACM, Jun. 2004, pp. 177-189, doi: 10.1145/990064.990087.

[36] A. Khoshgozaran and C. Shahabi, “Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy,” In Advances in Spatial and Temporal Databases, Springer Berlin Heidelberg, 2007, pp. 239-257, ISBN: 978-3-540-73539-7, doi: 10.1007/978-3-540-73540-3_14.

[37] M. Duckham and L. Kulik, “A formal model of obfuscation and negotiation for location privacy,” In Pervasive computing, Springer Berlin Heidelberg, 2005, pp. 152-170. ISBN: 978-3-540-26008-0, doi: 10.1007/11428572_10.

[38] C. Hung, L. Wei, and W. Peng, “Clustering Clues of Trajectories for Discovering Frequent Movement Behaviors,” In Behavior Computing, Springer London, 2012, pp. 179-196, ISBN: 978-1-4471-2968-4, doi: 10.1007/978-1-4471-2969-1_11.

[39] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, “Travel route recommendation using geotags in photo sharing sites,” In Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, Oct. 2010, pp. 579-588, doi: 10.1145/1871437.1871513.

[40] X. Long, L. Jin, and J. Joshi, “Exploring trajectory-driven local geographic topics in foursquare,” In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM, Sep. 2012, pp. 927-934, doi: 10.1145/2370216.2370423.

[41] Y. Lou, C. Zhang, Y. Zheng, and X. Xie, “Map-matching for low-sampling-rate GPS trajectories,” In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, Nov. 2009, pp. 352-361, doi: 10.1145/1653771.1653820.

[42] M. Baratchi, “Finding frequently visited paths: dealing with the uncertainty of spatio-temporal mobility data,” In Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on, Apr. 2013, pp. 479-484. IEEE, doi: 10.1109/ISSNIP.2013.6529837.

[43] H. Gowrisankar and S. Nittel, “Reducing uncertainty in location prediction of moving objects in road networks,” In Proceedings of the 2nd International Conference on Geographic Information Science, Sep. 2002.

[44] C. Schlenoff, R. Madhavan, and S. Balakirsky, “An approach to predicting the location of moving objects during on-road navigation,” In Proceedings of IJCAI-03-Workshop on Issues in Designing Physical Agents for Dynamic Real-Time Environments: World modeling, planning, learning, ancommunicating, Aug. 2003, pp. 71–79.

[45] J. Lee, W. C., & Krumm, “Trajectory Preprocessing,” In Computing with spatial trajectories, Springer New York, 2011, pp. 3-33, ISBN: 978-1-4614-1628-9, doi: 10.1007/978-1-4614-1629-6_1.

[46] M. Wachowicz and A. Ligtenberg, “Characterising the next generation of mobile applications through a privacy-aware geographic knowledge discovery process,” In Mobility, Data Mining and Privacy, Springer Berlin Heidelberg, 2008, pp. 39-72, ISBN: 978-3-540-75176-2, doi: 10.1007/978-3-540-75177-9_3.

A Package for the Homogenisation of Climate Data Using Geostatistical Simulation

Júlio Caineta, Sara Ribeiro, Roberto Henriques and Ana Cristina Costa
NOVA IMS
Universidade Nova de Lisboa
Lisboa, Portugal
Email: {jcaineta, sribeiro, roberto, ccosta}@novaims.unl.pt

***Abstract*—Climate data homogenisation is of major importance in monitoring climate change, the validation of weather forecasting, general circulation and regional atmospheric models, modelling of erosion, drought monitoring, among other studies of hydrological and environmental impacts. The gsimcli package is a work in progress project based on a geostatistical homogenisation method, aiming to make its application easier and more straightforward. It is expected that this contribution will help technicians, researchers and other professionals, to detect and to correct irregularities in climate data.**

***Index Terms*—Climate data; homogenisation; geostatistics; software.**

I. Introduction

Homogenisation of climate data is a very relevant subject since these data are required as an input in a wide range of studies, such as atmospheric modelling, weather forecasting, climate change monitoring, or hydrological and environmental projects. Often, climate data series include non-natural irregularities which have to be detected and removed prior to their use, otherwise it would generate biased and erroneous results.

In the last two decades, many methods have been developed to identify and remove these inhomogeneities [1][2][3][4]. One of those is based on a geostatistical simulation technique, direct sequential simulation (DSS), proposed by Soares [5], in which local probability density functions (PDFs) are calculated at candidate monitoring stations, using spatial and temporal neighbouring observations, and then are used for detection of inhomogeneities [6]. This approach has been previously applied to detect inhomogeneities in four precipitation series (wet day count) from a network with 66 monitoring stations located in the southern region of Portugal [7]. That study revealed promising results and the potential advantages of geostatistical techniques for inhomogeneities detection in climate time series.

This package is a product of a project (GSIMCLI – Geostatistical simulation with local distributions for the homogenisation and interpolation of climate data (PTDC/GEO-MET/4026/2012)) that aims to improve and develop that geostatistical homogenisation method, deploying its algorithms into a computer program.

The first studies of the method required a lot of time and interaction from its users. That happened mainly for two reasons: i) climate data may come from different sources and in different formats that have to be parsed, and ii) the method makes use of an already existent program, which has its own input and output formats. Handling different file formats among the algorithm's steps may require several transformations, back and forth.

This short paper introduces the method itself, describes the software development and its structure, illustrates an example of its usage, and finally lists some of the features and improvements that are expected in the near future.

II. Brief theoretical reference

A brief theoretical introduction to the related topics will be presented in this section.

A. Climate data homogenisation

The homogenisation of long meteorological time series is of extraordinary interest to the scientific community. The precise quantification of the variability of observed meteorological parameters is essential for many purposes. However, long instrumental records are rarely homogeneous because their values are dictated not only by change in climate but also by non-climatic factors. Irregularities such as relocation of weather stations or changes in measuring instruments, introduces discontinuities in time series, which may lead to data not being representative of real climate change. That may therefore skew the studies' results [1][6].

B. Geostatistical simulation

In geostatistics, it is common to refer to simulation as a stochastic process, opposed to estimation which is regarded as a deterministic process. Besides correlating different samples of a given variable, geostatistics adds their spatial structure to the equation. Therefore, geostatistical simulation is used to reproduce the spatial distribution and uncertainty of variables of different resources in Earth sciences.

One of the geostatistical simulation methods that has been widely used in different contexts (e.g., oil and gas resources, air and water pollutants) is the DSS. One of its main advantages is not requiring the transformation of the original variable, while honouring both the variable's covariance model and histogram.

C. Geostatistical simulation for the homogenisation of climate data

A geostatistical approach, using DSS, was proposed by Costa et al. [7] for inhomogeneities detection and correction. The DSS algorithm is used to calculate the local PDF at a candidate station's location, using spatial and temporal observations, only from nearby reference stations, without taking into account the candidate's data. Afterwards, the local PDF from each instant in time (e.g., year) is used to verify the existence of irregularities. A breakpoint is identified whenever the interval of a specified probability p (e.g., 0.95), centred in the local PDF, does not contain the observed (real) value of the candidate station [6]. In Figure 1, the orange areas illustrate the identification of a breakpoint, i.e., the values lying in the orange areas will be classified as inhomogeneous. In practice, the local PDFs are provided by the histograms of simulated maps. If irregularities are detected in a candidate series, the time series can be adjusted by replacing the inhomogeneous records with the mean, or median, of the PDF(s) calculated at the candidate station's location for the inhomogeneous period(s).

This technique accounts for the joint spatial and temporal correlation between observations, and gives greater weight to the nearest stations, both in spatial and correlation terms.

The final goal of the GSIMCLI project is to deliver a complete tool for the homogenisation of climate data, after investigating a new method based on DSS with local distributions [9]. It should result in a procedure that is appropriate for those situations in which the monitoring stations are located in extensive areas with different climatic characteristics, and it should be extensible to situations in which the PDF of the candidate station is different from its neighbours' PDF, which occurs, for example, due to local trends induced by local physiographic features.

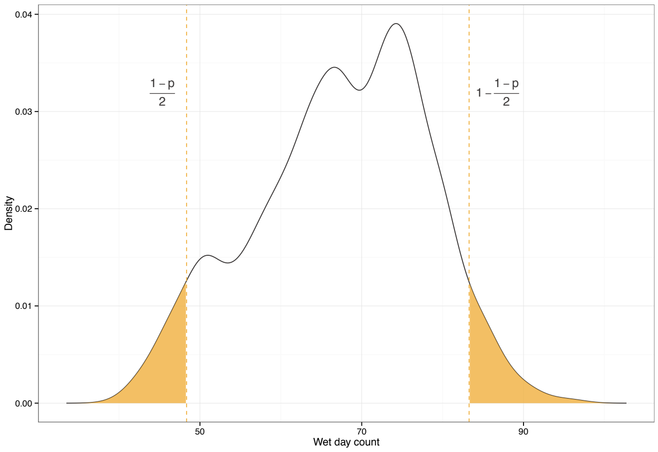


Fig. 1. Example of a probability density function originated from the simulation of the variable wet day count [8].

III. The gsimcli program

With the attention drawn into homogenisation methods by the HOME project (COST Action ES0601) [10] and the promising results attained in previous studies [6], the geostatistical approach became an interesting research subject, leading to this computer program.

The homogenisation method referred in Section II-C has been deployed into a computer program, which simplifies its application and assessment. In this section, we will elaborate on the program's details.

A. Software development

The gsimcli computer application (available at http://iled.github.io/gsimcli) is being developed under the object oriented paradigm, with the programming language Python (see http://www.python.org), a cross-platform, open source and general purpose language widely used in the scientific community. Its development is divided in a set of packages described below. The Python language is a good tool for prototyping (it is fast to code and easy to read) and for tasks that are not computationally demanding, still, it has a wide range of libraries that enhance its performance and usability.

The DSS algorithm is currently implemented as a black box element in the algorithm – the binary is launched as a standalone application – and its code is not presently part of this project.

B. Packages

This computer application can be seen as a collection of scripts, which have been divided into four packages. Each package encloses a different number of modules, then each module contains a set of related functions.

1) Parsers: One of the main purposes of the proposed program is minimising the required pre-processing and transformations needed during the homogenisation process, from the user point of view, i.e., the program should manage and transform all the data files automatically. Thus, a set of modules were developed to handle different file types, including reading, writing and the conversion between them:

costhome files in the format established for the benchmark dataset developed in the HOME project for the comparison of homogenisation algorithms.
dss DSS parameters file.
gsimcli parameters file of the gsimcli application.
shapefile files generated in a Geographical Information System (GIS) environment.
spreadsheet typical spreadsheet file (e.g., comma separated values).

2) Tools: The definition and handling of objects and the numerical calculations are mainly developed in these modules. There are algorithms implemented to: deal with objects that follow the specification given in the widely used geostatistical library GSLIB [11]; detect and identify breakpoints in time

ISBN: 978-1-61208-383-4

series (the main goal of this program); control parameters; and also to calculate performance metrics for the homogenisation process, i.e., centred root-mean-square error (CRMSE).

3) Launchers: This package operates the main processes execution: DSS and the homogenisation process as a whole.

The DSS related module takes advantage of the fact that different realisations are independent between them, to launch multiple simulations at the same time, in order to achieve a form of parallel execution. In this way, it is possible to use multiple cores, thus reducing the overall processing time.

The module that controls the homogenisation process includes options to make it run in time series separated in decades, and also in time series belonging to different networks (these functions are grouped with the term "batch"). With such functions, all the file handling is automated and user's perceive the homogenisation as a single process, instead of a repetitive and time-consuming sequence to process different input and output data.

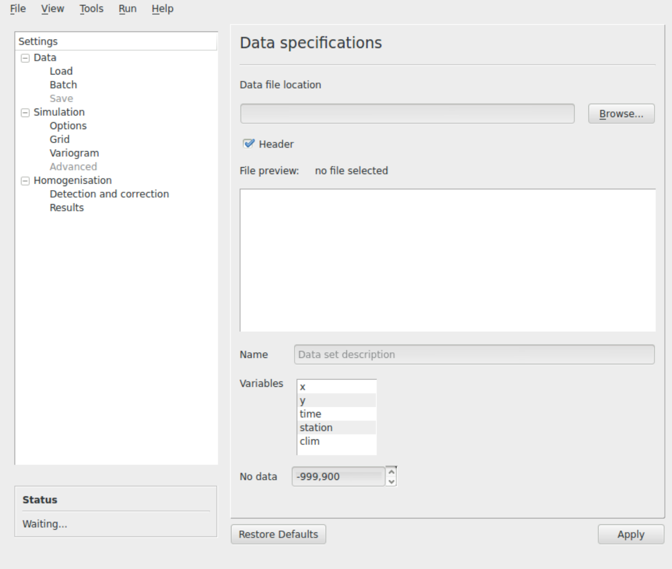
4) Interface: The program offers a graphical user interface (GUI) that was designed to be easy and intuitive to use, having a lot of common structures seen in other programs.

Basically, and at the current point of development, it is a settings pane with three groups of parameters: data, simulation (DSS) and homogenisation. They are organised in the order that the overall process will run and that should match a natural workflow (Figure 2).

5) Documentation: The program development also accounted for its documentation. There is an application programming interface (API) and a user manual for the GUI, both available at http://gsimcli.readthedocs.org.

C. Usage

As stated before, the program usage should be direct and effortless. A set of default and recommended settings is provided, which will help users to start using this application.



Data specifications

Data file location

☑ Header

File preview: no file selected

Name: Data set description

Variables: x, y, time, station, clim

No data: -999,900

Status

Waiting...

Fig. 2. Graphical user interface overview.

If using the default values, a common usage should go as follows:

1) Browse the data file location or, if processing multiple networks and/or decades, browse the data folder location.
2) Confirm the variables order and the place holder number for missing data.
3) Browse the DSS binary file.
4) Set up the simulation grid (if processing multiple networks, the grid details must be included in a spreadsheet file).
5) Provide the theoretical variogram model (if processing multiple decades, the variogram must be included in a spreadsheet file).
6) The settings for the detection and correction if inhomogeneities are also given by default, but it should be simple to try different values.
7) Browse the results file and directory.
8) Run gsimcli.

After that, the program will show the progress of the process. The necessary computational time highly depends on the computer specifications (e.g., frequency and number of cores), but also depends on the data set (number of candidate stations) and on the simulation parameters (e.g., grid size, maximum number of nodes to search for).

The final outcome is a spreadsheet file containing the complete homogenised data set, it will show the previous values of the homogenised samples, as well as a summary table indicating the number of detections and missing data per candidate station.

D. Enhancing the case study

When the geostatistical approach for the homogenisation of climate data was first studied, only 4 candidate stations out of 10 were selected (those 10 had been previously classified as inhomogeneous among a total of 66 stations) [7]. The process was slow and laborious, it required a considerable amount of user interaction and, for that reason, it was not practical to assess a desirable number of candidate stations. Nonetheless, that study revealed promising results and the potential advantages of geostatistical techniques for inhomogeneities detection in climate time series.

The implementation of all the required steps into computer scripts made it feasible to extend the same case study, with the number of candidate stations being increased to 10 (the referred inhomogeneous stations) [8]. Also, it is now practicable to analyse the method sensitivity to any parameter.

E. Performance assessment

The algorithms overall performance can be assessed in two aspects: the computational performance while running the algorithm; and the results of the homogenisation process. Both these perspectives were investigated and considered during the software development.

The computational performance has been an important factor in the design and implementation of the algorithms, both in processing time and required system memory. For example, as already mentioned, running multiple instances of DSS is a way to increase computational efficiency, while the division of time series in decades should help to reduce memory consumption.

In terms of quality of homogenisation, the performance has been measured in two ways: homogenising the southern Portugal case study and comparing with the results obtained in a previous study [7]; and measuring the obtained CRMSE in the homogenisation of some of the benchmark network of the HOME project, against other results presented by the same project [10]. In the first case, the results were promising and consistent with what had been obtained in the mentioned study [8]. In the latter, the results were considerably worse than those obtained by other homogenisation algorithms, given the small size of the monitoring networks and their low spatial dependence, thus highlighting the importance of the implementation of the DSS with local distributions [12].

IV. Conclusion and future work

The resulting computational application reproduces an existing and tested homogenisation method based on a geostatistical simulation technique, while having the advantage of the entire process being conducted in a seamless and practical manner, requiring less user interaction. This is highly beneficial to researchers: it is easier to investigate the influence of any parameter, and it allows the addition of new methods in a given step of the overall process, enhancing the research and development of new techniques and knowledge.

In the near future, it is planned to extend the GUI to include more options to the user, as well as to provide a graphical interface for other operations related to the homogenisation process (e.g., variography, conversion between file types, calculation of performance metrics).

Acknowledgement

The authors gratefully acknowledge the financial support of ”Fundação para a Ciência e Tecnologia” (FCT), Portugal, through the research project PTDC/GEO-MET/4026/2012 (”GSIMCLI - Geostatistical simulation with local distributions for the homogenisation and interpolation of climate data”).

References

[1] H. Caussinus and O. Mestre, “Detection and correction of artificial shifts in climate series,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 53, no. 3, pp. 405–425, Aug. 2004.

[2] A. T. DeGaetano, “Attributes of several methods for detecting discontinuities in mean temperature series,” *Journal of Climate*, vol. 19, no. 5, pp. 838–853, 2006.

[3] P. Domonkos, “Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods,” *Theoretical and Applied Climatology*, vol. 105, no. 3, pp. 455–467, 2011.

[4] T. Szentimrey, “Multiple analysis of series for homogenization (MASH),” in *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*, ser. WMO-TD No. 962, WCDMP No. 41, Budapest, Hungary, 1999, pp. 27–46.

[5] A. Soares, “Direct sequential simulation and cosimulation,” *Mathematical Geology*, vol. 33, no. 8, pp. 911–926, 2001.

[6] A. C. Costa and A. Soares, “Homogenization of climate data: review and new perspectives using geostatistics,” *Mathematical geosciences*, vol. 41, no. 3, pp. 291–305, Nov. 2009.

[7] A. C. Costa, J. Negreiros, and A. Soares, “Identification of inhomogeneities in precipitation time series using stochastic simulation,” in *geoENV VI - Geostatistics for Environmental Applications*, A. Soares, M. J. Pereira, and R. Dimitrakopoulos, Eds. Netherlands: Springer, 2008, pp. 275–282.

[8] J. Caineta, S. Ribeiro, A. C. Costa, R. Henriques, and A. Soares, “Inhomogeneities detection in annual precipitation time series in Portugal using direct sequential simulation,” in *Geophysical Research Abstracts*, ser. EGU General Assembly Conference Abstracts, vol. 16, no. EGU2014-7849, Vienna, Austria, 27 Apr.–2 May 2014, p. 7849.

[9] A. Horta and A. Soares, “Direct Sequential Cosimulation with Joint Probability Distributions,” *Mathematical Geosciences*, vol. 42, no. 3, pp. 269–292, Feb. 2010.

[10] V. K. C. Venema, O. Mestre, E. Aguilar, I. Auer, J. a. Guijarro, P. Domonkos, G. Vertacnik, T. Szentimrey, P. Stepanek, P. Zahradnicek, J. Viarre, G. Müller-Westermeier, M. Lakatos, C. N. Williams, M. J. Menne, R. Lindau, D. Rasol, E. Rustemeier, K. Kolokythas, T. Marinova, L. Andresen, F. Acquaotta, S. Fratianni, S. Cheval, M. Klancar, M. Brunetti, C. Gruber, M. Prohom Duran, T. Likso, P. Esteban, and T. Brandsma, “Benchmarking homogenization algorithms for monthly data,” *Climate of the Past*, vol. 8, no. 1, pp. 89–115, Jan. 2012.

[11] C. V. Deutsch and A. G. Journel, *GSLIB: Geostatistical Software Library*, ser. Applied Geostatistics Series. Oxford University Press, 1998.

[12] J. Caineta, S. Ribeiro, R. Henriques, A. Soares, and A. C. Costa, “Benchmarking a geostatistical procedure for the homogenisation of annual precipitation series,” in *Geophysical Research Abstracts*, ser. EGU General Assembly Conference Abstracts, vol. 16, no. EGU2014-7605, Vienna, Austria, 27 Apr.–2 May 2014, p. 7605.

Towards the Use of Crowdsourced Volunteered Meteorological Data for Forest Fire Monitoring

Shay Sosko and Sagi Dalyot
Mapping and Geo-Information Dept.
The Technion
Haifa, Israel
sosko@tx.tecnion.ac.il, dalyot@tx.technion.ac.il

***Abstract*— Static geosensor networks equipped with variety of sensors are used for collecting and transmitting physical data from their surrounding environment. These geosensor networks are used for monitoring potential disaster areas for early warning systems. Weather stations are an example of such static geosensor networks used for the collection of meteorological data, as temperature, humidity, pressure, wind direction and speed, used, for example, as input data for forest fire prediction and behavior systems. Nevertheless, static geosensor networks show some limitations, in terms of insufficient coverage, low density, nondynamical, limited power source and budget constraints. In recent years, the use of crowdsourced Volunteered Geographic Information (VGI) has emerged as a working methodology for retrieving real-time data in disaster areas; thus, allowing real-time and continuous aggregation of data and production of information. Using VGI methodologies, having the capacity of collecting dynamic real-time data can reduce deficiencies related to the use of static geosensor networks. This study is aimed at proving the feasibility of using VGI data as a reliable data source for reducing weather geosensor network limitations along with estimating the reliability and accuracy of crowdsourced data. This is done by analyzing the observations made by smartphones to prove their capacity. An example is presented to emphasize the potential of using crowdsourced VGI to densify and enhance static geosensor networks.**

Keywords-VGI; crowdsourcing; geosensor netowrk; densification; forest fires.

I. INTRODUCTION

Forest fires are vital natural phenomenon that maintains the health of ecosystems in reproducing, renewing and extending the diversity of species. Still, forest fires can also be a destructive natural disaster having devastating impact on the ecological, biological and social environment. Fast detection of fire and short arrival time of the fire brigades are key elements that can make the difference between small-scale disaster and mass casualty incident. Knowing in real-time crucial physical components that affect the spread and extent of forest fires enable the emergency agencies to act faster, and to be better prepared in case of forest fires. Because of the necessity for receiving - in real time - data from the area, early warning systems were developed.

A. *Geoseneor Networks for Early Warning Systems*

Early warning systems for disaster situations are usually based on static geosensor networks that monitor a predetermined stationary area for providing hazardous warning in cases of fires, floods, earthquakes and more. These geosensor networks are comprised of physical sensors aimed at warning of potential risk for disaster occurrence. Those sensors collect physical data from their environment and have the ability of transmitting it. Nevertheless, static geosensor networks show some limitations, in terms of insufficient coverage, low density, nondynamical, limited power source and budget constraints.

By using crowdsourced VGI working methodology, i.e., collecting data from volunteers in the area, limitations related to the use of static geosensor network can be reduced. For example, coverage and density of the network can be expanded using real-time user generated observations (negating the cost limitation), which if required, contribute to a more reliable data interpolation for acquiring information from unreachable areas. Moreover, the location of the reports varies in time, hence making the network dynamic and flexible. In addition, disabled sensors (e.g., damaged, blinded, or without power source) can be replaced with real-time data generated voluntarily by nearby citizens.

B. *VGI for Disaster Situations*

In recent years, the use of Volunteered Geographic Information (VGI) has emerged as a working methodology for retrieving real-time data in disaster areas. VGI is a subset of crowdsourcing paradigm, a working premise in which user-generated information is gathered and shared by individuals who participate voluntarily in a specific task. The reliability of information is usually derived by the number of volunteers in the area [1][2]. Geo-related crowdsourcing working paradigm that combines sensing and communication technologies, enables virtually everybody to collect data about the immediate environment - with almost no effort and expertise; thus, allowing real-time and continuous aggregation of data and production of information [2][3]. VGI is an effective method for data collection that can be used for expanding the variety of data sources and enhancing the resolution of sensor readings and reports, thus enriching and augmenting information derived from static sensors, especially when the nature of the geographic phenomena is dynamic [4].

In a disaster situation, geospatial data and tools have an essential part in many aspects of disaster management (planning, response, recovery and more). Using crowdsourcing tools, geo-tagged data can be collected by different means, allowing relief organizations to react better in case of emergencies and disasters [5]. It is widely acknowledged that real-time geospatial data are an essential source of information for all aspects of disaster management [6].

Generally, relying on user-generated data obtained in real-time from individuals nearby the interest area can augment the knowledge gap that might exist due to lack of accurate and updated data that is a result of inadequate sensor network deployment. Thus, enabling a more complete understanding of the disaster area. As in forest fires monitoring: fusing the crowdsourced sensor-data in real-time, and thus densifying existing static sensor-networks, have the potential to increase the likelihood to derive knowledge in respect to what is happening in real-time (i.e., now) – and where exactly. Also, to make a reliable and accurate assessment as to what will happen next, and consequently - how to respond.

This research is aimed toward improving and augmenting geosensor network deficiencies by using crowdsourced VG data. This paper presents preliminary examination of this research paradigm by analyzing meteorological data collected solely via smartphones (without the need of additional devices). The research is focused on determining the reliability and accuracy of the collected data when compared to authoritative meteorological geosensor networks. This will serve as a preliminary stage towards the feasibility of using the crowdsourcing paradigm for such scenarios.

The remainder of this paper is organized as followed: section II reviews related work in the field of VGI, section III introduces the methodologies used in this study, section IV describes the experiments, followed by analysis and evaluation of the results, which are detailed in section V. Section VI presents an example of the potential this paradigm have, and section VII concludes this paper and presents future work.

II. Related Work

Recent studies have proved that the public is collaborating in sharing and collecting information [7][8][9]. Furthermore, in cases of emergencies and disasters, the public's motivation for data collection is even bigger [10]. Various different VGI based solutions and platforms have emerged for emergency and disaster response. Ushahidi, for example, is a crowdsourcing platform used for the creation of crisis maps on the basis of integrating data from multiple sources and devices, as in the Haitian earthquake, Japanese tsunami, Kenyan post-election violence crisis [3]. OpenStreetMap (OSM) is a platform for creating GI by volunteering participants, which edit the map data by uploading GPS tracks, interpreting aerial imagery, out-of-copyright maps - and more [11]. This enables to use the platform for disaster situations characterized by the need for up-to-date maps (used in the 2010 Haitian earthquake) [3]. VGI can be used for flood damage estimation [12], radiation monitoring [13] or environmental sensing. Common Scent is an example of a crowdsourced based platform for environmental sensing that provides near real-time air quality data based on physical sensors mounted on bicycles [14]. Virtually, VGI enables to reduce the dependency on experts, while relying on the fact that accurate data is collected via diverse sources, consequently producing reliable information. The contribution of VGI for disaster situations is widely implemented in various different applications, assisting in better managing, controlling and recovering from such events, such as fires, earthquakes, violence riots, and also for environmental awareness, such as monitoring air quality, radiation level and other different hazards [5][15][16].

III. Methodology

A. Required Data for Forest Fire Model

The relevant crowdsourced data that is collected for fulfilling the objectives of this study can be divided to meteorological data and auxiliary data. The meteorological data are comprised of: (1) ambient temperature, and (2) relative humidity. These are the cardinal input data for fire spreading simulators (which generate information regarding ongoing fire: perimeter, growth, spreading rate and direction), and for fire danger rating systems (used for estimating the potential threat of a fire to ignite and spread over a large area) [17].

Auxiliary data assist in analyzing the reliability of meteorological data by describing the environmental conditions affecting the collection device. Since here data are collected via portable devices, auxiliary data are comprised of: (1) illumination – how exposed is the device to direct sunlight, (2) proximity - for detecting possible close by exterior disturbances, (3) battery properties – in terms of heating and current usage, and (4) GPS – acquiring the geographic position of the meteorological measurements. The readings of the first three give information that is relevant for the consistency of metrological data, since they might bias them and affect their reliability.

B. Data Collection Platform

Since crowdsourcing data collection relies on random individuals situated nearby the interest area, hence most probably portable devices will be used that have the capability to collect both the meteorological and auxiliary data and transmit it in real-time (via the internet). The more common and widespread the data collection platform is, the probability of citizens participating in gathering data increases, therefore enhancing the overall data – and hence information - accuracy, density and reliability. After examination of possible data collection platforms suitable for this purpose, the device that was chosen is the Samsung Galaxy S4 (SG4) model GT-I9500 (Samsung having a market share of 25% [18]). This model contains sensors required for the collection of all the aforementioned data types (such that no additional devices are needed). The application used for collecting and managing the data is 'WeatherSignal', which satisfies the collection requirements. The SG4 ambient temperature/relative humidity sensor model is SHTC1, manufactured by 'Sensirion'. The sensor is calibrated by the manufacturer for each device it is installed in before use in a controlled environment. The official accuracy of the sensor is depicted in Table I [19].

C. Reference Data

Analyzing the credibility of the crowdsourced collected data, it is necessary to have reliable and accurate sets of reference measurements of the same meteorological parameters; these will be used for comparison and evaluation purposes. For this, data from the Israel Meteorological Service (IMS) weather stations, which comply with the World Meteorological Organization (WMO) standards – also in terms of accuracy (depicted in Table II) [20] - are used.

IV. Field Experiments

To determine the accuracy of the SG4 ambient temperature and relative humidity measurements, data were collected in three different scenarios, which differ by the collection duration and location (environmental conditions), to enable a more qualitative assessment of collected data and measuring conditions.

In the first scenario, the aim was to verify that the measurements' accuracy is compatible with the official manufacturer accuracy (shown in Table I), and also in non-laboratory conditions only, i.e., field conditions as well. Therefore, data were collected continuously for a long duration (about 20 hours), while the SG4 was positioned statically nearby an IMS station in a shadowed place to eliminate the heating impact of exposure to direct sun light.

The second scenario was composed of a series of short duration measurements: four different times of day for approximately 1 hour, and nearby two alternative IMS stations. This scenario aims at determining the accuracy of the SG4 readings over short periods without having the need for data post-processing.

The SG4 measurements might be biased due to environmental conditions, which affect the measuring sensors. Therefore, it is essential to determine when do the SG4 readings are accurate without having to rely on external reference data. This is determined in the third scenario, which is composed from a series of five measuring sessions, where in each the SG4 was exposed for a short period to direct sunlight, which resulted in erroneous readings. Subsequently, the SG4 was moved from the sunlight to a shaded place until accurate measurements were obtained in comparison to the IMS data (with confidence level of 95%).

TABLE I. SHTC1 Sensor Official Accuracy.

Data Accuracy	**Meteorological Parameter**			
	Relative humidity (%)		***Ambient temperature (°C)***	
	Range	***Accuracy***	***Range***	***Accuracy***
SHTC1	20-80	4.5	5-60	0.4
	<20 or >80	7.5	<5 or >60	1.2

TABLE II. IMS Ambient Temperature and Relative Humidity Accuracy.

Data Accuracy	**Meteorological Parameter**			
	Relative humidity (%)		***Ambient temperature (°C)***	
	Range	***Accuracy***	***Range***	***Accuracy***
IMS	<50	3-5	<40	0.1-0.2
	>50	3	>40	0.3

V. Experimental Results and Evaluation

A. First Scenario

The results of the first scenario are displayed in Table III. The ambient temperature mean difference between the SG4 and the IMS measurements is 1.1°C, with Standard Deviation (SD) of 1.4°C. Assuming the data is normally distributed, the estimation interval of the mean with probability of 95% is 1.1°C ± 0.2°C. The relative humidity mean difference is 6.6%, with SD of 4%. The relative humidity confidence interval with probability of 95%, assuming the data is normally distributed, is 6.6% ± 0.7%.

Possible improvement of the measurements' accuracy can be achieved by eliminating outliers, which are caused by the surrounding environmental conditions affecting the sensor readings. Several outlier detection methods were implemented (IQR, boxplots, mean and SD methods). The outlier detection methods identified 6 to 9 outliers for the ambient temperature measurements, and 1 to 2 outliers for the relative humidity measurements; both, out of 134 observations, such that approximately 95% are considered as accurate and reliable. The mean and SD method was chosen to be used for eliminating outliers, since it was found to be more sensitive to outliers than the other methods.

After removing the outliers, the data were analyzed again, and the results are presented in Table IV. The ambient temperature mean difference was reduced to 0.8°C with SD of 0.7°C and confidence level of ±0.1°C. The relative humidity measurements were slightly reduced but remained of the same scale as prior to the outlier removal since not many outliers existed.

B. Second Scenario

The ambient temperature mean residual varies from minimal value of 0.3°C with SD of 0.1°C and similar confidence interval, to maximal value of 1.4°C with SD of 0.3°C and similar confidence interval, as can be seen in Table V. The average of overall mean residuals is 1°C, which is similar to the values obtained in the first scenario before the

TABLE III. Results of First Scenario: Measurements In Respect To Reference Data.

Statistical Analysis of Residuals	**Meteorological Parameter**	
	Ambient temperature (°C)	***Relative humidity (%)***
Mean	1.1	6.6
Standard Error	0.1	0.3
Standard Deviation	1.4	4
Minimal Residual	0	0.2
Maximal Residual	9.2	20.7
Count	134	134
Confidence Level (95.0%)	0.2	0.7

TABLE IV. FIRST SCENARIO RESULTS AFTER OUTLIER REMOVAL.

Statistical Analysis of Residuals	**Meteorological Parameter**	
	Ambient temperature (°C)	***Relative humidity (%)***
Mean	0.8	6.4
Standard Error	0.1	0.3
Standard Deviation	0.7	3.7
Minimal Residual	0	0.2
Maximal Residual	3.1	15.4
Count	126	132
Confidence Level (95.0%)	0.1	0.6

outlier removal. The relative humidity means residuals range from 1% with similar SD and confidence interval of 1%, to 14% with SD and confidence interval of 1%.

The mean residuals in 6 out of 8 temperature measurements were better or similar to the data obtained in the first scenario before removing outliers. In the relative humidity measurements, 5 out of 8 mean residuals were smaller than the first scenario data after removing outliers. According to the analysis results obtained for the two different scenarios in controlled conditions, it is clear that the SG4 can be used for acquiring accurate and reliable meteorological data, even without eliminating outliers, since it serves with good quality measures and low bias of measurements.

C. Third Scenario

The aim of this scenario was to determine the environmental conditions in which the SG4 measurements are accurate and reliable; this is aimed to allow not to use external reference data. Since sensors calibration time (needed for obtaining reliable results) are not constant and cannot be predetermined, four parameters were defined, which are calculated dynamically from the SG4 measurements, that help determine when the sensor readings are stable (not biased) and accurate, hence sensor is calibrated: (1) gradient (of data); (2) SD; (3) number of observations, and (4) illumination. Using statistical analysis, threshold values for the aforementioned parameters were determined, depicted in Table VI; based on these parameters a calibration algorithm was developed, depicted in Figure 1. Implementing the algorithm makes it possible to identify when the sensor readings are biased and/or erroneous, while correctly detect when readings are reliable - without having to use any external reference data.

Validation of the algorithm is depicted in Figure 2, showing a comparison of the IMS temperature data (reference) with VGI readings. It can be noted that the ambient temperature readings from both data sources are similar at the calibration point, characterized in the graph by horizontal slope of both indicators (SD and gradient of data), that was determined automatically by the algorithm.

VI. POTENTIAL OF VGI FOR REAL-TIME WEATHER DATA

The potential of using crowdsourcing for the collection of weather data is illustrated by the use of the crowdsourced weather map created by 'Weather Signal' (WS). WS is an application used for voluntarily collecting weather data (along with other available sensor data) from sensors mounted on portable devices (such as smartphones). The map depicted in Figure 3 demonstrates the high density VG data have in respect to the static network which is comprised of weather stations that are part of the IMS network used by KKL-JNF (Keren Kayemeth LeIsrael-Jewish National Fund) to monitor meteorological data nearby forests, thus on the augmentation potential of using crowdsourced weather data. It is clear that the VGI ambient temperature readings are filling the gaps in areas having no coverage of weather stations, densifying the impact zones, which are sparsely

TABLE VI. CALIBRATION THRESHOLD VALUES.

Calibration Threshold	**Calibration Parameters**			
	No. of Observations	***SD***	***Gradient***	***Illumination***
Ambient Temperature	30	0.5 (°C)	5%	<50,000(Lux)
Relative Humidity	30	1(%)	5%	<50,000(Lux)

TABLE V. RESULTS OF THE SECOND SCENARIO: MEASUREMENTS IN RESPECT TO REFERENCE DATA.

Meteorological Parameter	**Time of Measurement**	***GMT+3 [01:00-02:00]***		***GMT+3 [08:00-09:00]***		***GMT+3 [13:00-14:00]***		***GMT+3 [19:00-20:00]***	
	IMS Station Name	***Oil refinery***	***University***	***Oil refinery***	***University***	***Oil refinery***	***University***	***Oil refinery***	***University***
Ambient Temperature Residuals [°C]	Mean	0.3	0.9	1.4	0.9	1.2	1.5	1.1	1.1
	Standard Error	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.1
	Standard Deviation	0.1	0.1	0.4	0.3	0.5	0.3	0.2	0.1
	Minimal Residual	0.2	0.7	0.8	0.6	0.5	1.2	0.9	1
	Maximal Residual	0.4	1	1.9	1.6	1.8	1.8	1.3	1.3
	Confidence Level (95.0%)	0.1	0.1	0.4	0.3	0.5	0.2	0.1	0.1
Relative Humidity Residuals [%]	Mean	7	1	14	10	1	3	4	2
	Standard Error	1	1	1	1	1	1	1	1
	Standard Deviation	1	1	1	2	1	1	1	1
	Minimal Residual	7	0	12	8	0	2	1	1
	Maximal Residual	7	1	15	13	2	5	5	2
	Confidence Level (95.0%)	1	1	1	1	1	1	1	1

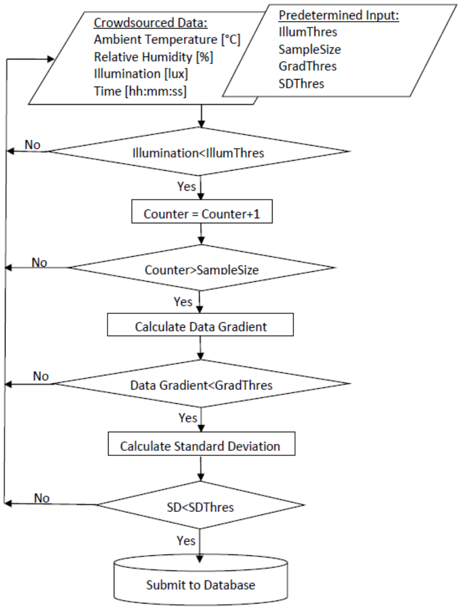


Figure 1. Calibration algorithm workflow.

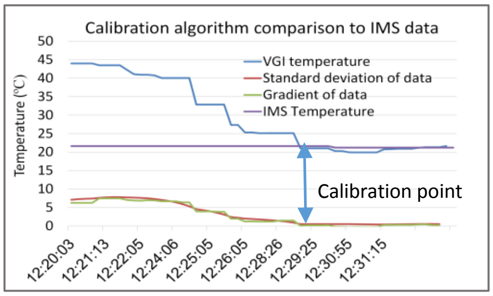


Figure 2. Algorithm indicators with IMS and VGI ambient temperature measurements.

represented by a single static IMS station. Moreover, it can be noticed that using crowdsourced data, areas currently not having any IMS station in their vicinity (south, center) now have several readings, which greatly contribute to a better assessment of physical conditions.

The crowdsourced weather data, downloaded from WS database for a specific date and time (22/3/2014, 13:00-14:00 GMT+2), were processed using spatial and attributional queries for eliminating measurements with incomplete data or characterized as indoor, which are irrelevant for this analysis. The algorithm used for this process was implemented using ArcGIS model-builder. Figure 4 depicts Kriging interpolation of the VGI ambient temperature readings, with and without existing IMS data for that specific date and time. Inspecting the interpolation results, it is clear that they are continuous and similar in values - VG data in respect to reference data (IMS) with no anomalies detected – supporting the fact that VGI measurements are reliable. More importantly, as depicted in the lower image, it is clear that some physical conditions are revealed and made clear and local (mainly in the center) thanks to the data densification – a direct result of using VGI.

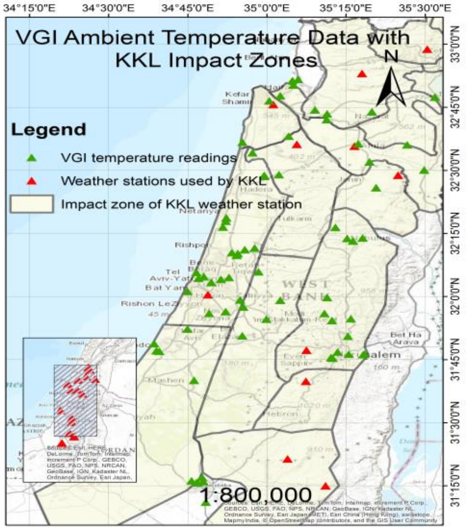


Figure 3. Map of VGI ambient temperature readings superpositioned with IMS weather stations used by KKL and their impact zones.

VII. Conclusions and Outlook

The feasibility of collecting reliable and accurate VGI metrological observations (ambient temperature and relative humidity) via smartphones (SG4 in this case study) was verified. Consequently, as presented in the previous section, the use of crowdsourced volunteered meteorological measurements is made possible, enhancing the input weather data collected from the surrounding. Although the accuracy of the 'SHTC1' sensor might decrease due to environmental conditions, the algorithm developed for this research is capable to detect such situations, thus having the possibility to filter out erroneous observations. This proves that with relatively small post-processing, and without having the need to use reference data to analyze the correctness of the data, the collection device can function as a reliable and accurate 'dynamic geosensor station' that serves as supplementary data source that is external and independent to the static network. Though other devices - other than the SG4 - should be analyzed, first inspection made with available WS data showed that alternative portable devices are also reliable.

The next phase of this research will be focused on the collection and analysis of crowdsourced VG data from wider areas, and on the development of a fusion algorithm designed for observations from crowdsourced VGI and IMS. These are aimed for densifying the static geosensor network, while the output of such process will augment and improve the input data required for forest fire models.

Acknowledgment

The authors would like to thank IMS and KKL-JNF for delivering with information regarding the position and impact zones of IMS. Also, to the OpenSignal team, that

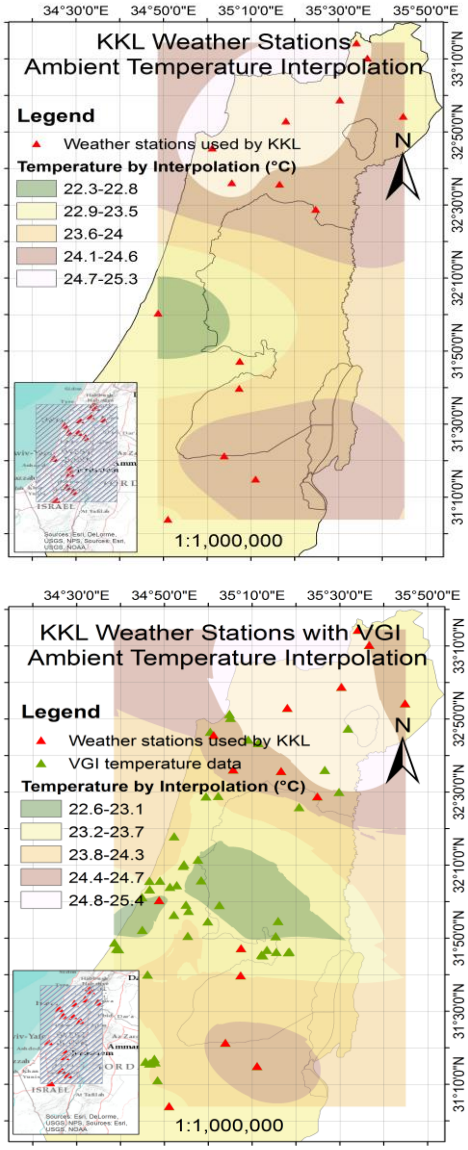


Figure 4. Kriging interpolation map of IMS stations ambient temperature measurements (top), and together with VGI readings (bottom).

delivered with valuable information and assistance regarding their WS web service and data.

REFERENCES

[1] E. Estellés-Arolas and F. González-Ladrón-de-Guevara, "Towards an integrated crowdsourcing definition," Journal of Information science, vol. 38, no. 2, Apr. 2012, pp. 189-200, , doi:10.1177/0165551512437638.

[2] M. Goodchild and J. Alan Glennon, "Crowdsourcing geographic information for disaster response: a research frontier," International Journal of Digital Earth, vol. 3, no. 3, Apr. 2010, pp. 231-241, doi: 10.1080/17538941003759255

[3] M. Zook, M. Graham, T. Shelton, and S. Gorman, "Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake," World Medical & Health Policy, vol. 2, no.2, Jul. 2010, pp. 7-33, doi:10.2202/1948-4682.1069.

[4] J. W. S. Liu, E. T. H. Chu, and P. H. Tsai, "Fusing human sensor and physical sensor data," 5th IEEE International Conference on Service-Oriented Computing and Applications (SOCA 2012), IEEE, Dec. 2012, pp. 1-5, doi:10.1109/SOCA.2012.6449429.

[5] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the Crowdsourcing Power of Social Media for Disaster Relief," IEEE Intelligent Systems, vol. 26, no. 3, May 2011, pp. 10–14, doi:10.1109/MIS.2011.52.

[6] National Research Council (US). Committee on Planning for Catastrophe: a Blueprint for Improving Geospatial Data, Tools, and Infrastructure. Successful Response Starts with a Map: Improving Geospatial Support for Disaster Management. National Academies Press, 2007.

[7] G. Bugs, C. Granell, O. Fonts, J. Huerta, and M. Painho,"An assessment of Public Participation GIS and Web 2.0 technologies in urban planning practice in Canela, Brazil," Cities, vol. 27, no.3, Jun. 2010, pp. 172-181, doi:10.1016/j.cities.2009.11.008.

[8] D. J. Coleman, Y. Georgiadou, and J. Labonte, "Volunteered Geographic Information: the nature and motivation of produsers," International Journal of Spatial Data Infrastructures Research, vol. 4, no.1, Jan. 2009, pp. 332-358, doi:10.2902/1725-0463.2009.04.art16.

[9] E. Hand, "Citizen Science: People power," Nature, vol. 466, Aug. 2010, pp. 685–687, doi:10.1038/466685a.

[10] K. Starbird, "Digital volunteerism during disaster: Crowdsourcing information processing," Workshop on crowdsourcing and human computations, CHI, May 2011.

[11] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," IEEE Pervasive Computing, vol. 7 no. 4, Oct. 2008, pp. 12-18, doi:10.1109/MPRV.2008.80.

[12] K. Poser and D. Dransch, "Volunteered geographic information for disaster management with application to rapid flood damage estimation," Geomatica, vol. 64, no. 1, 2010, pp. 89–98.

[13] A. Saenz, "Japan's Nuclear Woes Give Rise to Crowd-Sourced Radiation Maps in ASIA and US,". [Online]. Available from: http://singularityhub.com/2011/03/24/japans-nuclear-woes-give-rise-to-crowd-sourced-radiation-maps-in-asia-and-us, [accessed: 2014-11-17].

[14] B. Resch, R. Britter, C. Outram, R. Xiaji, and C. Ratti, "Standardised Geo-sensor Webs for Integrated Urban Air Quality Monitoring", Environmental Monitoring, 2011, pp. 513-528, ISBN: 978-953-307-724-6.

[15] D. Bird, K. Haynes, M. Ling, and J. O'Brien, "The use of crowd sourcing for gathering information about natural disasters," Risk Frontiers newsletter, vol. 11, no. 2, Dec. 2011.

[16] K. O. Asante, J. P. Verdin, M. P. Crane, S. A. Toker, and J. Rowland, "Spatial Data Infrastructures in Management of Natural Disasters," Research and theory in advancing spatial data infrastructure concepts, 2007, pp. 279-293, ESRI Press.

[17] A. Patricia and B. Butler, "Fuels Management-How to Measure Success", Rocky Mountain Research Station Conference Proceedings, RMRS-P-41, Mar. 2006, pp. 201-212.

[18] IDC, "Smartphone OS Market Share, Q2 2014". [Online]. Availble from: http://www.idc.com/prodserv/smartphone-os-market-share.jsp, [accessed: 2014-12-17].

[19] J. Robinson, "The SHTC1: Inside the chip that powers WeatherSignal,". [Online]. Availble from: http://opensignal.com/blog/2013/06/19/the-shtc1-inside-the-chip-that-powers-weathersignal, [accessed: 2014-12-17].

[20] World Meteorological Organization, "Guide to Meteorological Instruments and Methods of Observation", Seventh edition, 2008.