Data Mining for Forecasting Fog Events and Comparing Geographical Sites

Designing a novel method for predictive models portability

Gaetano Zazzaro, Gianpaolo Romano
Italian Aerospace Research Centre, CIRA
Capua (CE), Italy
email: g.zazzaro@cira.it
email: g.romano@cira.it

Paola Mercogliano
Euro-Mediterranean Center on Climate Change, CMCC,
Italian Aerospace Research Centre, CIRA
Capua (CE), Italy
email: paola.mercogliano@cmcc.it

***Abstract*— Fog represents high impact atmospherical phenomena especially for aviation. Low visibility conditions severely affect air traffic operations especially during the landing and take-off phases and thereby reducing the capacity of an airport. In particular, in 2001 the Linate Airport in Milan was hit by a disaster, the deadliest air disaster to ever occur in Italian aviation history, due to un-forecasted thick fog. For this reason, improvement of fog monitoring and forecast tool is a challenge topic for the aviation community. Moreover, forecasting fog is an important issue for air traffic safety because adverse visibility conditions represent one of the major causes of traffic delay and of the economic loss associated with such phenomena. In such context, the present work illustrates a Data Mining application for the fog forecasting on a short time range (1 hour) on Linate airport. Indeed two predictive models have been trained using an historical dataset of 18 years of fog observations including many meteorological parameters collected in the Synop message. These models have been made up by applying BayesNet and Neural Network algorithms. The performances evaluation highlights that the complete model shows 90% of instances correctly predicted. Moreover, in order to discover whether predictive models trained on Milan can also be used for forecasting fog events on other geographic sites, a new method to characterize fog events and compare different airport areas is described. Thus, a novel metric is defined, aimed at comparing different sites. This metric is based on the Euclidean distance between performance vectors that are also here defined. Thanks to this metric, we can determine whether a new set of fog observations is compatible or not with Linate fog observations and whether, formally, the predictive models are portable to the new site. Furthermore, we are able to group geographical locations that can be also many kilometers distance away. This work represents a first design step to define the comparative metric. It has been carried on according to the standard process (CRISP-DM) for Knowledge Discovery in Database Process.**

Keywords-Data Mining; Forecast Fog; Bayesian Networks, Artificial Neural Networks; Inductive Decision Trees; Model Portability; CRISP-DM.

I. INTRODUCTION

This paper is an extended version of the conference paper [1]. With respect to the conference version, in this paper we expand the description of the methodology adopted for the creation of fog predictive models on Linate Airport in Milan, including more details of Data Understanding and statistical exploratory charts. Moreover, a new descriptive model of fog events is explained and the design of an its innovative use is introduced, aimed at comparing different geographical sites; in addition, we discuss the portability of predictive models to other sites that are similar (or compatible, with a meaning that will be detailed in this paper) with the Milan Linate site.

The effort spent by aeronautic research on this topic is due to the importance to reduce the impact on the different flight phase (e.g., taxing, landing, take off) on the atmospherical phenomena. This requirement is obtained improving the current capability to forecast (on different time range) adverse weather condition.

Fog forecast and its characterization represent a challenging topic due to the local condition causing this phenomena. Moreover, low visibility conditions severely affect air traffic operations especially during the landing and take-off phases and thereby reduce the capacity of an airport. This leads to the built-up of a wave of delayed flights in case demand exceeds the reduced capacity, which is especially critical at major hubs, such as, for Italy, Milan Linate during peak times. Since these hubs are central nodes in the air traffic network, the effect also spreads causing the event to be of much more than just local importance. Indeed the occurrence of low ceilings and/or poor visibility conditions restricting the flow of air traffic into major airport terminals is one of the major causes of traffic delay and of the economic loss associated with such phenomena [2]. For these reasons, a fast forecasting tool is crucial to adequately manage the occurrence of these events and to mitigate their impact over the whole airport system. Consequently, it is important to deeply understand the process leading to the formation of fog and justifies the efforts made by meteorologists to forecast such events.

In this paper, a method for fog nowcasting (short-range forecasting of 1 hour) on Linate Airport in Milan based on Data Mining techniques is presented. Indeed Data Mining (DM) [3] – also called "Knowledge Discovery in Databases" – refers to the process of extraction or "Mining" useful knowledge from large amounts of data. DM draws upon ideas, such as sampling, estimation, and hypothesis testing from statistics and search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning.

DM can represent a useful analysis method for this complex meteorological phenomenon because it has the

ability to work with many data described by a high number of variables.

In order to obtain DM models for fog prediction, we used an historical dataset consisting of 164.352 meteorological SYNOP observations collected at Milan's Linate airport station from January 1996 until September 2014.

Knowledge Discovery in Database Process, that we carried on in order to predict fog events, has been conducted according to the standard process conceived from the Cross-Industry Standard Process for DM (CRISP-DM) [4]. CRISP-DM is structured in six steps (Figure 1).

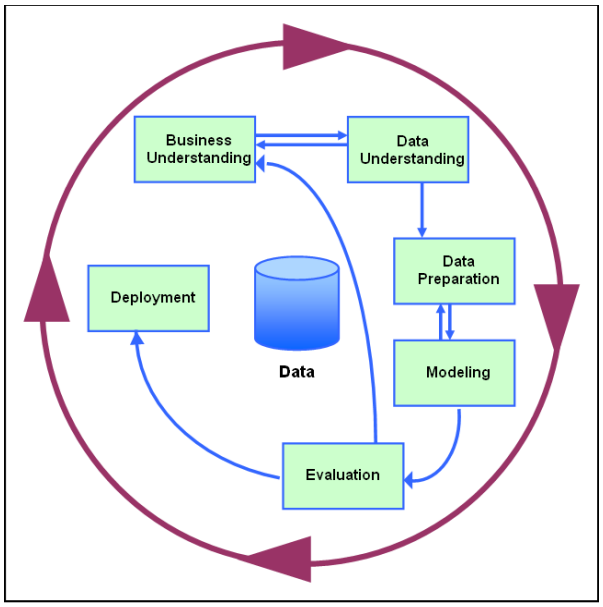

Figure 1. CRISP-DM Steps

Every step of the process has been supported by the validation of domain experts. In this work, we used the Weka tool (Version 3.8.0) (Waikato Environment for Knowledge Analysis) [5] to carry on DM analysis. In particular, we used the Weka Explorer interface to mine data by applying Bayesian Nets, Artificial Neural Networks, and Inductive Decision Trees algorithms.

A. *Structure of the paper*

The paper is organized by describing all the CRISP phases one by one. In Section II, the Business Understanding is carried on in order to understand the fog phenomenon and its development, to explore the state of the art from meteorological and DM points of view, and to fix DM goals. In Section III, we illustrate the data collection, the data sources, the variables and statistics of the attributes. In Section IV, we explain all the activities of the Data Preparation phase aimed at constructing the final datasets to be mined; in particular, the preprocessing phase and the hold-out method. In Section V, we provide details about the Modeling phase: from the identification of target functions to model testing. Moreover, we illustrate a descriptive model by which we are able to better understand the fog phenomenon and to define a measure of similarity between meteorological stations (or geographic locations). In Section VI, we present the Evaluation phase including the analysis of the misclassified fog events. In Section VII we design how to use the descriptive model and its future investigations ad applications. Finally, in Section VIII, we show our considerations.

II. BUSINESS UNDERSTANDING

This first step of the CRISP-DM process includes fixing of the business objectives, Data Mining goals and assess situation.

A. *Fog Formation and State of the Art of Fog Nowcasting*

Fog is basically a cloud of small water droplets near ground level and sufficiently dense to reduce horizontal visibility to less than 1 km (3281 feet). The word fog also may refer to clouds of smoke particles, ice particles, or mixtures of these components. Under similar conditions, but with visibility greater than 1000 m, the phenomenon is termed a mist or haze, depending on whether the obscurity is caused by water drops or solid particles. The formation of fog is due to the condensation of water vapor on condensation nuclei (non-gaseous solid particles) to form water droplets, near the ground. Fog usually develops when relative humidity is near 100% and when the air temperature and dew point temperature are close to each other or less than 4˚F (2.5˚C). When air reaches 100% of relative humidity, its dew point is said to be saturated and can thus hold no more water vapor. As a result, the water vapor condenses to form water droplets and fog. The formation of fog is a complex process involving highly non-linear interactions between surface and sub-surface processes, atmospheric radiation, turbulence and flows. Such interactions are not adequately described by the current operational Numerical Weather Prediction (NWP) [6], because the vertical and horizontal resolutions are larger than the corresponding fog scales [7] that are of the order of 1 km on the horizontal scale and up to few ten meters on the vertical scale. For these reasons these models [2] [8] [9] are unable to treat complex three-dimensional flows due to their poor representation of horizontal heterogeneities [7].

In order to overcome such limitations, dedicated NWP models have been implemented [10] in order to predict the formation of fog in regions of complex terrain and reach horizontal grid resolution of 1km or better. The disadvantage of such models lies on the computational costs required to run them [6]. For this reason, they can be applied only on small domains and on high speed computer [6].

Finally, the statistical methods [11] can overcome the above-mentioned problems but they require long time series of homogeneous data and they can be used only for specific locations for which the fog events can be correlated to the local conditions. In fact fog events can be triggered by different physical causes and their characteristic strongly depends on the specific geographical location [12].

Traditional data analysis techniques (including statistical and physical driven techniques) have been often faced with practical difficulties in meeting the challenges posed by new datasets including meteorological datasets (with a high number of records, variables, sources, etc.). DM techniques can represent useful analysis methods because they are able to

investigate different meteorological variables coming from numerous datasets. DM techniques provide a high level of prediction in terms of consistency and frequency of correct predictions.

Prediction is the most used DM task in meteorology domain. DM has been applied successfully to predict different weather elements like wind speed [13] [14], rainfall [15] [16], cloud [17] and temperature [18] [19].

DM description task is carried on in [20] and [21] by using Decision Trees and Bayesian Networks in order to create some fog local indices, based on the post-processing of meteorological variables. The same methods were used in [22] for creation of some basic neural network structures that were further adapted to local prediction models. This approach was implemented and tested in various conditions of major Australian airports.

The fog formation and its important parameters were identified based on collected historical dataset from the International Airport of Rio de Janeiro [23]. In [24] the authors describe three short-range fog-forecasting models by applying Bayesian Networks in order to predict fog events between 0-3 hours on Paris Charles De Gaulle airport.

The availability of a long time series data set (SYNOP data) together with the necessity to describe such phenomenon in a specific site (Milan's Linate airport), make the DM approach one of the best solutions in describing and short range forecasting fog phenomena.

B. Business Objectives and Data Mining Goals

The Business objective is to develop an algorithm, which is able to describe, and nowcast fog phenomenon over Milan's Linate Airport, using DM techniques and Synop data. In particular, the objective is to forecast a fog event on the time range of 1 hour, associating a prediction probability. Classification models will be trained in order to forecast fog events.

Of course, probabilities can be transferred into crisp event forecasts, but since developments in air traffic management systems point towards more and more automation and decision support, direct use of probabilities will be favored because it enables detailed cost benefit analysis for triggering decisions

III. DATA UNDERSTANDING

This step of the CRISP-DM includes the initial data collection, data description, data exploration, and the verification of data quality.

A. Data Collection

In order to build a predictive model using DM techniques for fog forecast, a historical dataset made up of fog observations and relevant meteorological parameters needs to be built.

Data have been collected from ECMWF MARS Archive [25] containing the surface Synoptic observations (SYNOP) provided by Linate meteorological station.

SYNOP observations are recorded every hour. A list of the meteorological variables used for DM and selected from the SYNOP message is reported in TABLE I.

B. Fog Event Description

Each fog event can be defined as a sequence of SYNOP records with a visibility attribute value less or equal than 1000 meters. Each record describes the weather conditions observed.

TABLE I. LIST OF METEOROLOGICAL VARIABLES (FEATURES)

#	Name	Description	Units
1	Date	Date of the observation	Date
2	Pressure	Force per unit area exerted against a surface by the weight of the air above that surface	Pa
3	three hour pressure change	Change of the pressure with respect to three hours ago	Pa
4	char pressure tendency	Coded values indicating how the pressure has changed during one hour	-
5	wind direction	Wind direction at 10 m	Deg
6	wind speed	Wind speed at 10 m	kn
7	Visibility	It represents the greatest distance at which a black object of suitable dimensions (located on the ground) can be seen and recognized when observed against the horizon sky during daylight or could be seen and recognized during the night if the general illumination were raised to the normal daylight level. Visibility values below 1 km can indicate the presence of fog	m
8	present weather	Coded values describing the weather phenomena present at the time of the observation. Values between 40-49 indicate the presence of fog	-
9	past weather1	Coded values describing weather phenomena occurring during the preceding hour	-
10	past weather2	Coded values describing weather phenomena occurring during the two preceding hours	-
11	cloud cover	Values between 0 and 8 indicating the fraction of the celestial dome covered by all clouds visible. It is estimated in eighths (okta) of sky covered by clouds. Clear sky is indicated with 0 okta, overcast with 8	okta
12	height of base of cloud	Height of bases of clouds above ground level	m
13	cloud type	Coded values reporting the type of cloud and the state of sky	-
14	Dewpoint	Temperature at which moist air saturated with respect to water at a given pressure has a saturation mixing ratio equal to the given mixing ratio (ratio between the mass of water vapour and the mass of dry air)	°C
15	Drybulb	Temperature of the air measured with a thermometer shielded to radiation and humidity	°C

Fog events are characterized by an initial and final SYNOP message: the first recording is the head of the event; the last one is the end of fog event; one or more persistences of YES are between the head and ending in the single fog event.

In Figure 2, two examples of fog events are reported: the first event lasts three hours and the second one lasts two hours (the second event has no persistence of YES because it lasts only two hours and the first hour is the head while the last one is the end).

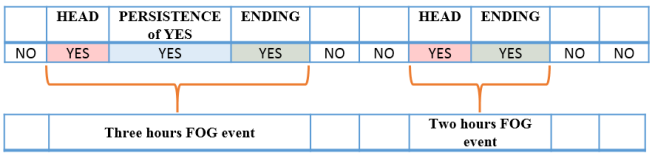

Figure 2. Sequences of recordings

C. Data Exploration

The collected dataset contains 164.352 instances belonging to the period from 1st January 1996 until 30th September 2014. Using the Weka's explorer interface [5] [29] we are easily able to view histograms for each attribute in TABLE I and plot matrices of different attribute combinations. Weka also displays basic statistics for each numeric attribute. In the following, some histograms are reported in order to investigate data and variables. For example, Figure 3 reports the number of instances of Dewpoint variable in the considered dataset. Dewpoint histogram presents a distribution similar to a Gaussian one.

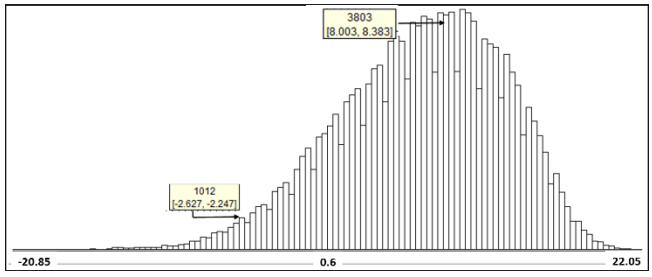

Figure 3. Histogram of instances by Dewpint attribute.

In TABLE II some basic statistics of Dewpoint attribute are reported.

TABLE II. STATISTICS OF DEWPOINT ATTRIBUTE

Statistic	Value
Minimum	-20.85
Maximum	22.05
Mean	8.001
StdDev	5.636
Missing	10741 (6.53%)
Distinct	375

The dewpoint temperature has a very low minimum value. This indicates that there are some outliers in the data set, which are removed in the next CRISP step.

Figure 4 reports the number of instances of Pressure variable in the dataset considered. Also Pressure histogram presents a distribution similar to a Gaussian one. In addition, table of basic statistics is reported in TABLE III.

IV. DATA PREPARATION

In order to obtain the final dataset that can be used in the modeling phase, data have been preprocessed to report them in a format usable by DM algorithms. In the original dataset there are 10676 missing records corresponding to the same number of missing hours. For these recordings, we have only date and time variables. The other attributes are all null. These missing records are removed from the original dataset, obtaining a new dataset with 153.676 instances.

TABLE III. STATISTICS OF PRESSURE ATTRIBUTE

Statistic	Value
Minimum	94950
Maximum	102750
Mean	100194.228
StdDev	882.571
Missing	51 (0%)
Distinct	669

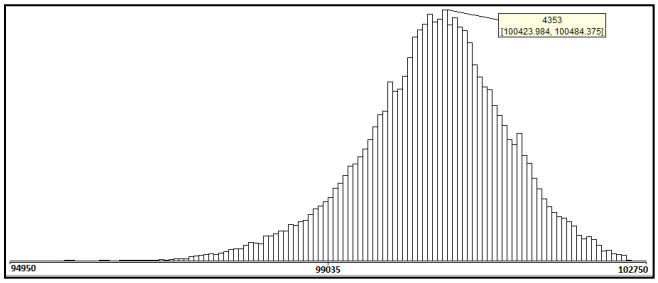

Figure 4. Histogram of instances by Pressure attribute

A. Variables Transformation and Target Class Creation

The meteorological parameters coded according to the World Meteorological Organization (WMO) code tables [26] have been converted from numeric to nominal type in order to report them in a format usable by DM algorithm. Such conversion is also required for a clearer reading of data and results. After the conversion, the target attribute has been identified according to the domain expert indications. Indeed the presence of fog is detected if visibility is less than or equal to 1 km [27].

Moreover, *Date* variable is splitted into two new attributes: *Month* and *Hour*.

The histogram of target class of Figure 5 shows how fog is a quite rare meteorological event on Linate airport: fog occurs about once every 53 events. Target class is unbalanced.

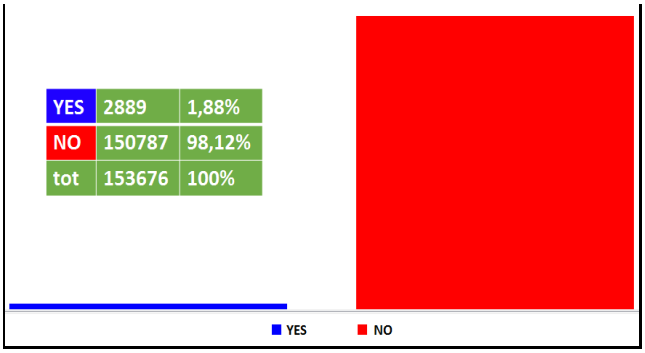

Figure 5. Histogram of instances by class target attribute

In order to visualize the distribution of FOG according to the variation of variables, the graph of Figure 6 shows that fog events, which are represented in blue color, occur mostly from October to March. In addition, from the histogram of instances

by Hour attribute (not reported), fog events occur in the early hours of the morning and in the late evening.

Figure 7 shows the scatter plot for Dewpoint and Drybulb variables with the line bisector. Fog events are in blue color. Since the points of the line bisector have Dewpoint=Drybulb, in the area close to the upper side of the line bisector where the fog events are mostly distributed, fog events have a small positive value of (drybulb–dewpoint).

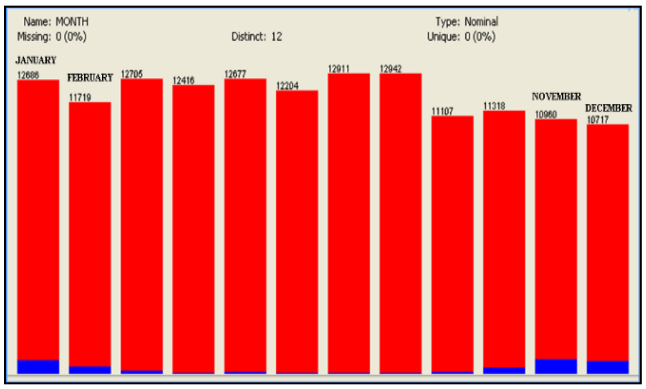

Figure 6. Histogram of instances by Month attribute, from Jan. to Dec.

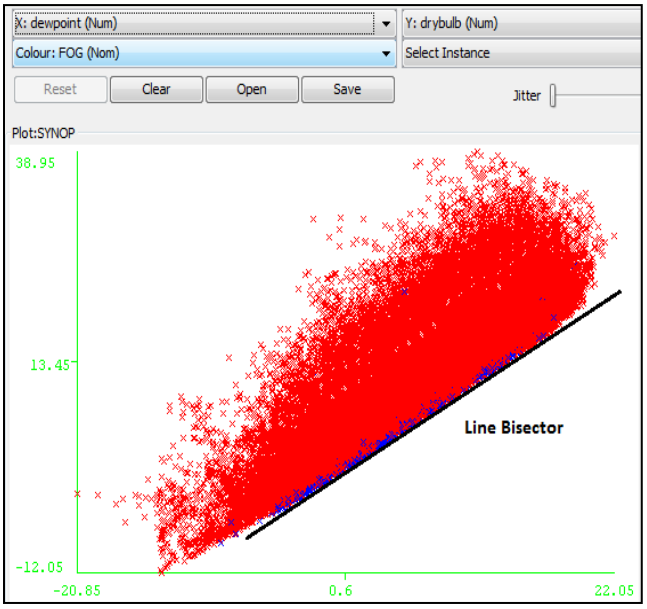

Figure 7. Scatter plot for Dewpoint and Drybulb variables with Line Bisector

In total, the Linate weather dataset has 17 (16+1) attributes, including FOG target class.

B. Hold-Out Method and Forecast Sets Preparation

For DM goal, we adopted the working strategy named hold-out method [28]. In this method, the original data with labeled examples is partitioned into two disjoint sets, called training and test sets, respectively. A classification model is induced from the training set and its performances are evaluated on the test set. The accuracy of the classifier can be estimated based on the accuracy of the induced model on the test set.

As test set we choose the records belonging to the last 13 months of meteorological observations, from 1st September 2013 until 30th September 2014. This test set is called 1_YEAR, it has 9314 full weather observations and it has roughly the same target class distribution of whole dataset.

FOG and 1_YEAR are useful to train and test a descriptive model, called *F-model*, as described in the next section.

V. MODELING

Modeling follows Data Preparation; in Modeling phase descriptive and forecast models are trained and tested.

A. The F-model

In order to better understand the fog phenomenon on Milan Linate airport, a descriptive Data Mining model is realized. This model, called *F-model*, takes into account the actual weather conditions and it is able to recognize whether the measured weather conditions are related to Milan fog events or not:

$$F: (feat_1, feat_2, \dots, feat_{14}, t_0) \rightarrow \{YES, NO\}_{t_0} \in FOG$$

Input variables are 14 (and not 16) because *Visibility* and *present weather* attributes are obviously removed (they are directly and closely correlated to FOG class).

Thus, F is a system of classification rules that has at least two main uses:

1. describe fog events at the Linate airport;
2. provide a "similarity metric" able to determine whether the fog meteorological conditions of an airport are similar to those at Milan Linate airport. Thanks to F we can determine whether a new set of fog observations is compatible or not with Linate fog observations.

FOG set is used to training the *F-model* while 1_YEAR set is used to test it. FOG set has got 152971 full weather observations, 1_YEAR has got 9314 records.

In the current section, *F-model* is described in order to understand data, explore fog events and recognize them at Linate airport; whilst in Section VII, a novel use of the descriptive model is suggested in order to compare fog events at different airports and to group geographic locations that can be also very distance away (from the spatial point of view).

The descripted *F-model* is an Ensemble model trained combining different classifiers by using a majority voting strategy [30] [31]. A combined classifier often shows better performance than individual classification models (component classifiers). An optimal set of classifiers is first selected and then combined by a specific fusion method, in order to obtain the *F-model*.

Furthermore, the improvement of the generalization performance of such ensemble model will be demonstrated experimentally (TABLE XI).

The *F-model* is 10-cross folds validated, and it is obtained by combining 4 Bayesian Nets (BN) [29] and 9 Inductive Decision Trees (IDT) [28] [32], by changing the values of the parameters of BayesNet and J48 Weka algorithms. The choice of BN and IDT is due to [20] [21].

The BayesNet algorithm of Weka allows to define such components by means of the following parameters:

- *searchAlgorithm* selects the method for searching network topology; we fixed it to K2.
- *Estimator* selects the algorithm for calculating the conditional probability tables. We chose the SimpleEstimator algorithm. This algorithm has a parameter, called A = alpha parameter, which sets the starting value for the calculation of conditional probability.

Four models have been produced, by changing the values of the following P parameter:

- P = maxNrOfParents parameter of K2 algorithm which sets the maximum value of the number of parents (maximum number of edge arrows) of each node in the net topology.

In particular, A is fixed to 0.5 and P ranges from 1 to 4 in order to obtain 4 Bayesian networks.

The J48 algorithm of Weka allows to generate a pruned or unpruned C4.5 decision tree. Its main parameter is N = "NumOfFolds", that determines the amount of data used for reduced-error pruning (one fold is used for pruning, the rest for growing the tree). Moreover, the "reducedErrorPruning" J48 parameter is fixed to "True" value (whether reduced-error pruning is used instead of C4.5 pruning).

By changing N from 2 to 9 and fixing one time "reduceErrorPruning" = "False", 9 Inductive Decision Trees are trained. TABLE IV summarizes the 13 trained models.

TABLE IV. 13 DESCRIPTIVE MODELS

BayesNet	J48	Total
P = 1, …, 4	N = 2, …, 9 +	
A = 0.5	reduceErrorPruning = "False"	
4 BN	9 IDT	13 models

As reported in Figure 4, target class FOG is imbalanced; the Weka filter SpreadSubsample under-samples the dataset in order to obtain the same number of FOG = "YES" and FOG = "NO" instances. This balancing technique is used in order to overcome the class unbalance problem. In particular, the balanced Training set has 5502 instances (2751 FOG="NO" + 2571 FOG="YES"). Another way to overcome the class unbalance problem is the cost-sensitive technique [21].

All obtained descriptive models of fog events have been compared and the achieved results have been evaluated by means of adequate performance metrics able to highlight the classifying ability with respect to the fog events and the no-fog events separately (e.g., confusion matrix, AUC) [28] [29].

Finally, the Ensemble *F-model* has the performances of TABLE V, TABLE VI, and TABLE VII, considering 10-cross folds validation as test mode:

TABLE V. *F-MODEL* EVALUATION

Total Number of Instances	5502	
Correctly Classified Instances	5367	(97.5463 %)
Incorrectly Classified Instances	135	(2.4537 %)

TABLE VI. *F-MODEL* DETAILED ACCURACY BY CLASS

=== Detailed Accuracy By Class ===				
Class	TP Rate	FP Rate	Precision	ROC Area
YES	0.976	0.025	0.975	0.975
NO	0.975	0.024	0.976	0.975

TABLE VII. CONFUSION MATRIX OF *F-MODEL*

Forecast		← Classified as	
YES	**NO**		
2684	67	**YES**	**Observed**
68	2683	**NO**	

F-model shows the performances on 1_YEAR Test Set included in TABLE VIII, TABLE IX, and TABLE X. FOG class target in 1_YEAR test set has, roughly, the original unbalanced distribution of the whole dataset.

TABLE VIII. *F-MODEL* EVALUATION ON *1_YEAR*

Total Number of Instances	9314	
Correctly Classified Instances	9198	(98.7546 %)
Incorrectly Classified Instances	116	(1.2454 %)

TABLE IX. *F-MODEL* DETAILED ACCURACY BY CLASS

=== Detailed Accuracy By Class ===				
Class	TP Rate	FP Rate	Precision	ROC Area
YES	0.977	0.012	0.535	0.983
NO	0.988	0.023	1	0.983

TABLE X. CONFUSION MATRIX OF *F-MODEL* ON *1_YEAR*

Forecast		← Classified as	
YES	**NO**		
130	3	**YES**	**Observed**
113	9068	**NO**	

The Ensemble model has better performance than its single components. The following TABLE XI shows some metrics of the components of *F-model* on 1_YEAR test set, including *F-model*.

TABLE XI. *PERFORMANCES OF THE COMPONENTS OF F-MODEL* ON *1_YEAR TEST SET*

#	Model	True Positive Rate (TPR)	True Negative Rate (TNR)	AUC
1	BN_1	0.887	0.992	0.981
2	BN_2	0.962	0.985	0.982
3	BN_3	0.97	0.989	0.989
4	BN_4	0.962	0.988	0.99
5	IDT_2	0.977	0.986	0.985
6	IDT_3	0.962	0.984	0.987
7	IDT_4	0.977	0.983	0.98
8	IDT_5	0.977	0.985	0.985
9	IDT_6	0.977	0.984	0.984
10	IDT_7	0.977	0.986	0.991
11	IDT_8	0.977	0.981	0.99
12	IDT_9	0.97	0.981	0.99
13	IDT_false	0.97	0.98	0.98
14	***F-model***	**0.977**	**0.988**	**0.983**

The best component model is the Inductive Decision Tree trained fixing N=7 (called IDT_7). In particular, IDT_7 has the highest TPR (0.977) and the highest TNR (0.986). The *F-model* exceeds IDT_7 because *F* has an higher TNR than IDT_7.

B. A and B Models Design

The DM models should be able to predict fog after one hour from the recording of the last available SYNOP data. So, in order to easily forecast fog events, a new dataset is released starting from the dataset available after Data Preparation step. Shifting upwards of a position the time series of FOG variable, we obtain a new target attribute (FOG+1) describing the condition of fog at time $t_{0+1hour}=t_1$, while the meteorological attributes remains at time t_0. Such elaboration allows getting a new training set, called FOG+1, and a new test set, called 1_YEAR+1.

The one-hour prediction model has to be able to recognize both the beginning and the end of a fog event. Therefore, two models have been trained, *A-model* and *B-model*:

3. *A-model* is used in order to predict the persistence of NO and the discontinuities from NO to YES (heads of new fog events).
4. *B-model* is used in order to predict the persistence of YES and the discontinuities from YES to NO (endings of fog events that are heads of NO-fog events), instead.

As a consequence, *A-model* is used when the occurring visibility (visibility at time t_0) is greater than 1000 m and *B-model* in the other cases. A summary of this criterion is reported in TABLE XII.

TABLE XII. RULE FOR MODELS APPLICATION

if visibility at time t_0 > 1000m then *A-model* else *B-model*

DM models are simple predictors for time series, where the prediction of outputs for time t_1 is based on the sequence of historical data observed at time t_0.

In order to obtain two predicting models (*A-model* and *B-model*), each one of two datasets (FOG+1 and 1_YEAR+1) has been splitted in two subsets.

*A*_FOG+1, *B*_FOG+1 aim at train the *A-model* and the *B-model*, respectively, *A*_1YEAR+1 is useful to evaluate the performances of *A-model*, while *B*_1YEAR+1 is useful to evaluate the performances of *B-model*. The next schema of Figure 8 summarizes the steps of the Data Preparation phase.

In particular, after the step I, detailed in Figure 7, the original dataset has been cleaned and splitted in two subsets; after the step II the label class of the two datasets has been upward shifted for one hour.

Finally, in order to obtain the training and the test sets for A-model we have selected FOG="NO", and to obtain the training and test sets for B-model we have selected FOG="YES".

Therefore, we have applied the rules of TABLE XIII and TABLE XIV.

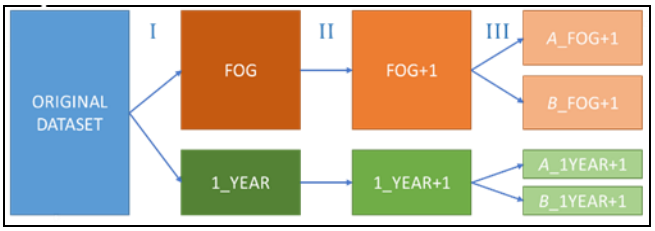

Figure 8. Forecast Sets Preparation Schema

TABLE XIII. *RULE FOR A_FOG+1 AND A_1YEAR+1 SETS*

FOG	FOG+1	
NO	NO	← Persistence of NO
NO	YES	← Head of fog event

TABLE XIV. *RULE FOR B_FOG+1 AND B_1YEAR+1 SETS*

FOG	FOG+1	
YES	YES	← Persistence of YES
YES	NO	← Ending of fog event

In addition, in order to overcome the class imbalance problem (Figure 4), the class labels of training sets have been under sampled, obtaining the same numbers of records with FOG+1="NO" and FOG+1="YES". However, the two test sets retain the original class target distributions. Finally, the Data Preparation produces the four datasets presented in TABLE XV, including their sizes.

TABLE XV. DATASETS ROLES AND DIMENSIONS

	A-model	***B-model***
Training	*A*_FOG+1 1380	*B*_FOG+1 1392
Test	*A*_1YEAR+1 9046 records	*B*_1YEAR+1 135 records

Starting from the two new datasets *A*_FOG+1 and *B*_FOG+1, we are able to train forecast models by using DM techniques. Indeed a forecast model is a function that takes into account the meteorological variables measured at time t_0 and computes a binary variable FOG+1 that indicates the presence or absence of fog at time t_1 and the respective probabilities.

In the next sections, the best obtained models are described but, for the sake of clarity, in our project many predictive models have been trained and only the performances of a Bayesian Net and an Artificial Neural Network are highly satisfactory for one-hour fog predictions on Linate airport database. The testing of the two 1-hour classification models show good performances, as follows.

C. The A-model

The *A-model* is a Bayesian Network classifier. It has been trained on the *A*_FOG+1 dataset (obtained from FOG+1 set using the instances tagged by FOG="NO"). For the sake of clarity, the training set *A*_FOG+1 is obtained by balancing the target class FOG+1, using the filter SpreadSubsample of Weka tool.

In this way, A-set presents 690 records tagged by FOG+1=”NO” and 690 records tagged by FOG+1=”YES”. The *A-model* is trained by using BayesNet Weka algorithm.

Fixing P=3 and A=0.25, the *A-model* performs on 10-fold cross-validation and it shows the performances included in TABLE XVI, TABLE XVII, and TABLE XVIII.

TABLE XVI. *A-MODEL* EVALUATION

Total Number of Instances	1380	
Correctly Classified Instances	1214	(87.971 %)
Incorrectly Classified Instances	166	(12.029 %)

TABLE XVII. *A-MODEL* DETAILED ACCURACY BY CLASS

=== Detailed Accuracy By Class ===				
Class	TP Rate	FP Rate	Precision	ROC Area
YES	0.884	0.125	0.876	0.932
NO	0.875	0.116	0.883	0.932

TABLE XVIII. CONFUSION MATRIX OF *A-MODEL*

Forecast		← Classified as	
YES	**NO**		
610	80	**YES**	**Observed**
86	604	**NO**	

TABLE XIX. *A-MODEL* EVALUATION ON *A_1YEAR+1*

Total Number of Instances	9046	
Correctly Classified Instances	8480	(93.7431 %)
Incorrectly Classified Instances	566	(6.2569 %)

TABLE XX. *A-MODEL* DETAILED ACCURACY BY CLASS

=== Detailed Accuracy By Class ===				
Class	TP Rate	FP Rate	Precision	ROC Area
YES	0.732	0.062	0.051	0.934
NO	0.938	0.268	0.999	0.934

TABLE XXI. CONFUSION MATRIX OF *A-MODEL* ON *A_1YEAR+1*

Forecast		← Classified as	
YES	**NO**		
30	11	**YES**	**Observed**
555	8450	**NO**	

The *A-model* shows the performances on *A*_1YEAR+1 Test Set included in TABLE XIX, TABLE XX, and TABLE XXI. This test analyzes the capability of the *A-model* to predict the persistence of the condition FOG=”NO” or the presence of the head of the fog events (FOG=”YES”).

D. The B-model

The B-classifier is an Artificial Neural Network (ANN) [28] trained on the balanced *B*_FOG+1 dataset (obtained from FOG+1 set using the instances tagged by FOG=”YES” and balancing the target class FOG+1 by using the Weka filter SpreadSubsample). In this way, *B*_FOG+1 presents 696 records tagged by FOG+1=”NO” and 696 records tagged by FOG+1=”YES”.

TABLE XXII. THE *B-MODEL* EVALUATION

Total Number of Instances	1392	
Correctly Classified Instances	1207	(86.71 %)
Incorrectly Classified Instances	185	(13.29%)

TABLE XXIII. THE *B-MODEL* DETAILED ACCURACY BY CLASS

=== Detailed Accuracy By Class ===				
Class	TP Rate	FP Rate	Precision	ROC Area
YES	0.888	0.154	0.852	0.891
NO	0.846	0.112	0.883	0.891

The *B-model* is trained by using the MultilayerPerceptron [29] algorithm of Weka tool, with 10 hidden layers (H=10) and N=1000 that is the number of epochs to train through. It performs on 10-fold cross-validation and it shows the performances included in TABLE XXII, TABLE XXIII, and in TABLE XXIV.

TABLE XXIV. CONFUSION MATRIX OF THE *B-MODEL*

Forecast		← Classified as	
YES	**NO**		
618	78	**YES**	**Observed**
107	589	**NO**	

B-model shows the performances on *B*_1YEAR+1 of TABLE XXV, TABLE XXVI, and TABLE XXVII.

TABLE XXV. THE *B-MODEL* EVALUATION ON *B_1YEAR+1*

Total Number of Instances	135	
Correctly Classified Instances	109	(80.74%)
Incorrectly Classified Instances	26	(19.259%)

TABLE XXVI. THE *B-MODEL* DETAILED ACCURACY BY CLASS

=== Detailed Accuracy By Class ===				
Class	TP Rate	FP Rate	Precision	ROC Area
YES	0.828	0.25	0.881	0.814
NO	0.75	0.172	0.614	0.814

TABLE XXVII. MATRIX OF THE *B-MODEL* ON *B_1YEAR+1*

Forecast		← Classified as	
YES	**NO**		
82	17	**YES**	**Observed**
9	27	**NO**	

This test analyzes the capability of the *B-model*, when the instances FOG=”YES” is present, to predict in the following hour the persistence of the condition FOG=”YES” or the presence of the end of the fog events.

VI. MODEL EVALUATION

Evaluation of the performance of a classification model is based on the number of test records correctly and incorrectly predicted by the model. Good results correspond to large

numbers along the main diagonal of the confusion matrix and small, ideally zero, off-diagonal elements.

The confusion matrix of the *A-model* on *A*_1YEAR+1 Test Set shows 555 records incorrectly classified as "YES" (TABLE XII), corresponding to 555 false positives instances (555 recordings without fog incorrectly predicted as heads of fog events).

TABLE XXVIII shows the distribution of such False Positives by Month attribute. The 74% of False Positive instances occur in [September, January].

TABLE XXVIII. DISTRIBUTION OF FALSE POSITIVES BY MONTH

# of records	Month	Total number of hours in the month
100	September 2013	720
49	October 2013	744
102	November 2013	720
99	December 2013	744
62	January 2014	744
12	February 2014	672
73	March 2014	744
18	April 2014	720
15	May 2014	744
5	June 2014	720
15	July 2014	744
5	August 2014	744
Tot=555		

Figure 9 shows the histogram of False Positives by Hour attribute. About 80% of False Positive instances occur in [00:00, 09:00] (range of Hour attribute).

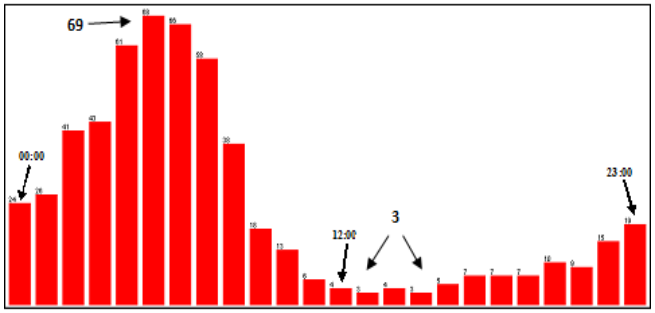

Figure 9. Histogram of false positives by Hour attribute

In addition, about 75% of False Positive instances occur when visibility ranges within [1200m, 4500m] (range of Visibility attribute). About 70% of False Positives occur when the 'Height of base of cloud' attribute is within [30m, 1000m] range and about 81% in [0kn, 7.77kn] range of 'Wind speed'. Therefore, False Positives have higher occurrence when these favorable meteorological conditions for fog presence are recorded, as low wind speed intensity and low cloud.

TABLE XXIX shows the distribution of False Positives by 'Present Weather' attribute.

The histograms and the statistic distributions prove that most of predicted false positives occur when the observed visibility conditions are below 5 km due to the presence of meteorological conditions that can reduce visibility (mist, drizzle, rain or fog). It has been considered that the present model considers only prediction of low visibility due to fog presence, while there are also other physical sources causing the reduction of the visibility. Therefore, even if these events are classified as false positives for fog event presence (because the observed visibility is greater than 1000 m), they correctly classify the events being physically characterized by low visibility conditions.

TABLE XXIX. FALSE POSITIVES DISTRIBUTION BY 'PRESENT WEATHER'

# of records	Present weather
56	Drizzle
17	Rain
9	Fog
305	Mist
136	No Meteors
25	Fog or Ice Fog
7	Patches
Tot=555	

Furthermore, most of the incorrectly predictions occur during months and hours often interested by fog events (autumn and winter seasons, night and early hours of the day), and during which a reduction of visibility conditions occur.

In conclusion, the *B-model* performs worse than the *A-model*. However, this evaluation does not worry us considering the significant increase of flight safety.

Anyway, considering the difficulty of the prediction of this atmospherical phenomenon results can be considered very promising for further investigation.

VII. HOW TO USE THE LINATE DESCRIPTIVE MODEL

In this section, a novel and alternative use of the *F-model* obtained in Section V is described.

A. Portability

A-model and *B-model*, built using Data Mining techniques applied to the weather data of Milan Linate airport site, are hardly able to predict fog events in other geographic locations without a meaningful loss of performance; in other words they are not "portable" to other different sites.

This inability has at least two explanations, the former in the meteorological field and the latter in the statistical field:

1. Meteorological inability: the geo-physical conditions of the various sites can be very different. Local classical forecast models, though they may have a similar analytical structure, are usually tailored to the geographic area using different known parameters.

2. Statistical inability: a model built with Data Mining techniques has a satisfactory accuracy only if it is tested on a set "compatible" with the set used for training; for example, it occurs when the test set is a random subset of the training set. Distributions of weather variables of data sets coming from different sites can be very "discordant". Obviously, "concordant" statistical distributions are a

necessary but not sufficient condition for model applicability in other different sites.

Getting a general purpose model able to predict fog events irrespective of the site being monitored is a very complex undertaking and may even fail if adequate methods are not taken into account and specific techniques are not applied.

A useful strategy for obtaining a "general purpose" model could be to consider a broader training set, obtained by merging the weather data sets of the various sites. Training such a model could approach too different situations and several attempts have led to an increase of the classification error.

In this section a novel strategy is described, which takes into account a single analytical model that exceeds the critical issues of points 1 and 2. A criterion is defined to determine whether the predictive model previously made at the Milan airport is portable or not on a different site, based on an automatic comparison among sites by defining a distance (a metric) by which to establish a "similarity" criterion among databases and then geographic locations.

In particular, this strategy has various corollaries, aimed at enhancing knowledge of the domain and meteorological phenomenon of fog.

B. Procedure Design

The algorithmic scheme developed for the prediction model portability is based on the idea that the predictive models of Milan Linate are also applicable to another site *S* if *S* is compatible (or similar) with the Milan site.

This paper describes the procedure for comparing a new geographic site *S* with the Milan Linate site using the descriptive model *F*.

The application and testing of the procedure is not here reported, but there is only the procedure design.

TABLE XXX. PERFORMANCE VECTOR MEASURES

1.	CCR	Correctly Classified instances Rate	0.987
2.	KS	Kappa Statistic	0.686
3.	TPR	True Positive Rate	0.977
4.	FPR	False Positive Rate	0.012
5.	PR	Precision	0.535
6.	MCC	Matthews Correlation Coefficient	0.718
7.	AUC	Area Under the ROC Curve	0.983
8.	PRCA	Area under Precision Recall Curve	0.471

The *F-model* has a performance vector *REFP* that summarizes its behavior on 1_YEAR test set. The chosen metrics [28] [29] [33] [34] [35] of *REFP* are in TABLE XXX, where also their values are reported. This *REFP* vector can be considered as a reference point for calculating the similarity measure of other geographical sites.

Considering now a new site S, the performance vector, *T_S_V*, derived from the *F-model* testing on the *S* set, is calculated. *S* is a set of meteorological observations recorded in the same time period of 1_YEAR test set, and its variables have been prepared in the same way as the Milan Linate weather variables.

The Euclidean distance, $ED(S)$, between *T_S_V* and *REFP*, can be considered as a similarity measure between S and Milan Linate. This measure of similarity, or compatibility, causes sorting between the various sites: the S_1 site is more compatible (or more similar) than the S_2 site, with respect to Milan Linate, (and you write $S_1 >_c S_2$), if $ED(S_1) < ED(S_2)$.

After finding the closest sites (more similar) to Milan Linate using the descriptive model *F*, you can try exporting predictive models *A-model* and *B-model* to these new geographical sites as well. If the tests are good, predictive models are portable on new geographic locations. Another alternative predictive model may be [24].

C. Compatibility Schema

Next TABLE XXXI summarizes the symbols used and useful for the Compatibility Schema.

TABLE XXXI. SYMBOLS OF COMPATIBILITY SCHEMA

Name	Description
F-model	A descriptive model trained on Milan Linate dataset [01/01/1996, 08/31/2013]
1_YEAR	Milan Test Set [09/01/2013, 09/30/2014]
REFP	The Performance Vector of *F*, with metrics calculated on 1_YEAR
S	A new geographical Site S is a dataset of weather observations
T_S	*S* Test Set [01/09/2013, 30/09/2014]
T_S_V	The Performance Vector, with metrics calculated on *T_S*
$ED(S)$	The Euclidean distance between *T_S_V* and *REFP*

The Compatibility Schema is the procedure useful to easily understand if a new geographical site *S* is "compatible" with Milan Linate. In the case of compatibility, *A-model* and *B-model* may also be applied to predict fog events on the new *S* site. For the sake of clarity, obviously, a small value of $ED(S)$ is a necessary but not sufficient condition for model portability.

The scheme in TABLE XXXII aims at comparing S_1 and S_2 sites, calculating their distances with Milan Linate separately. $S_1 >_c S_2$ only means that S_1 is closer to Milan than S_2, but S_1 may not be close enough to Milan. It is important to analyze an adequate number of sites, to determine a proximity threshold (that is an Euclidean distance), under which predictive models are portable.

Figure 10 summarizes the compatibility schema for predictive models portability, extended to *n* geographical sites.

The descriptive *F-model* is evaluated on 1_YEAR (=T_MIL) and on the other test sets $T\_S_1, , T\_S_2, \ldots, , T\_S_n$; the $n+1$ descriptive performance vectors are calculated, where the first one is *REFP*; all the n Euclidean distances $(ED(S_1), ED(S_2), \ldots, ED(S_n))$ are measured; and finally these distances are sorted, in order to discover the sites more similar (or compatible) to Milan and to study the portability of the predictive models to these sites.

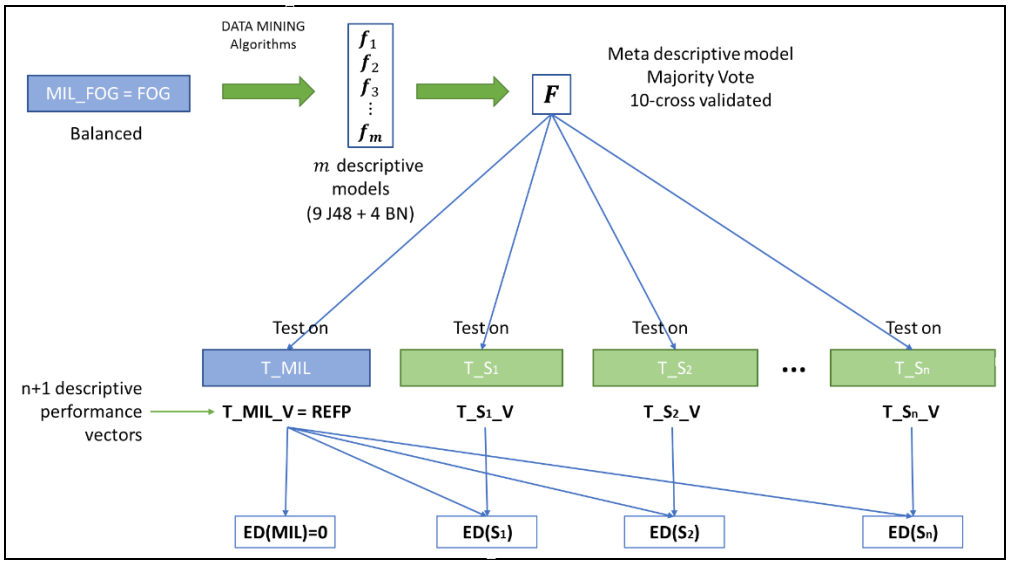

Figure 10. Compatilibility Schema for *n* Sites

TABLE XXXII. COMPATIBILITY SCHEMA

#	Step	Note
1	Data Preparation	Preparation of S_1 and S_2 datasets, including outliers detection and cleaning, and Target Variable creation. Selecting of test sets T_S_1 and T_S_2. This Data Preparation steps are the same steps of Milan Linate preparation steps (FOG and 1_YEAR)
2	Modeling	Training of the Linate descriptive *F-model*
3	Evaluation	Calculation of performance vectors *REFP*, $T\_S_1\_V$ and $T\_S_2\_V$, by considering 8 metrics: $(CCR, KS, TPR, FPR, PR, MCC, AUC, PRCA)$
4	Evaluation	Calculation of Euclidean distances *d*: $ED(S_1) = d(T\_S_1\_V, REFP)$ $ED(S_2) = d(T\_S_2\_V, REFP)$
5	Evaluation	Comparison of distances $ED(S_1) < ED(S_2)$ and then $S_1 >_c S_2$
6	Testing	Testing of *A-model* and *B-model* on T_S_1 to decide if models are portable to S_1

D. Future Investigations

Of course, the next steps will be dedicated to testing the compatibility procedure between geographical sites, calculating performance vectors and their distances to Milan, and also validating predictive models on the other sites. The predictive models can be the *A-model* and *B-model* of this paper but also the Bayesian Networks illustrated in [24].

The Euclidean distance among the performance vectors obtained on many geographic sites thanks to the *F* function, can be used as a measure of distance in a clustering algorithm (for example, k-means clustering), useful for obtaining homogeneous groupings of geographical sites. These sites can be colored on a map depending on whether they belong to a cluster, obtaining, for example, a new layer in a GIS.

VIII. CONCLUSIONS

This paper reports the main features of a statistical tool implemented for the forecast, in a very rapid time, the occurrence of low visibility events (or fog events) over the airport area. This method is essentially based on the use of an historical time series of SYNOP data available over Milan Linate airport and on the Data Mining techniques. SYNOP are a meteorological data message available in many airports, therefore the method can potentially be extended easily to different other airports. Two different classifiers have been trained in order to obtain two models that together are able to predict fog events on 1 hour time range. In order to reach this aim, the Data Understanding, Data Preparation, Modeling and Evaluation phases of CRISP-DM have been carried out.

Data Understanding phase includes the collection, description and exploration of data used for DM. Data Preparation phase allowed to elaborate data in order to obtain the dataset to be used for Modeling phase. In the Modeling phase, two different forecasting models (*A-model*, *B-model*) have been produced by applying BayesNet and Neural Network algorithms. Preliminary results show that the two models encourage the forecast of fog events on 1-hour time range. The *A-model* presents a percentage of correct classified instances of 93.74% and a percentage of true positive rate of about 73.2% corresponding to heads of fog events correctly predicted. Additionally the *B-model* presents a percentage of correct classified instances of 80.74% and a percentage of true positive rate of 75% corresponding to ends of fog events correctly predicted. Furthermore, both models have a very high percentage of correct classification of persistences of FOG="NO" and FOG="YES".

In addition to *A* and *B* models, this work has also illustrated how a descriptive model *F* was trained. *F* is an Ensemble meta-model, based on Bayesian Networks and Inductive Decision Trees. It is useful for better

understanding the fog phenomenon and as a preliminary step to define new method for fog forecast. It can be mainly used to apply the Milan predictive models in other sites. Thus, *F* is useful to define a new similarity measure between different geographical sites, useful for determining the portability of the predictive models in future applications. In this way, the geographic locations space can be clustered, in order to identify the sites that are more compatible or more similar to the Linate site.

In addition, future investigations could quantify the performances for detecting sharp transients, i.e., change of status from no-fog to fog and vice versa.

ACKNOWLEDGMENT

The authors would express their gratitude for funding part of this work in equal parts to the SESAR programme (www.sesarju.eu) funded by the European Union, Eurocontrol and its industrial members and to Selex ES GmbH. This work have contributed to the design of an integrated Ground Weather Monitoring System (GWMS) in SESAR project 15.04.09.c lead by Selex ES. Moreover, the authors would also mention the project TECVOL II founded by the Italian PRORA where the upgrade of the tool has been developed.

REFERENCES

[1] G. Zazzaro, P. Mercogliano, and G. Romano, "Bayesian and Artificial Neural Network for Nowcasting Rare Fog Events," in The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA), 2017, pp. 84-90.

[2] T. Bergot, D. Carrer, J. Noilhan, and P. Bougeault, "Improved Site-Specific Numerical Prediction of Fog and Low Clouds: A Feasibility Study," Weather and Forecasting 20, 627–646, 2005.

[3] P. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," Pearson Addison Wesley, 2005.

[4] P. Chapman et al., "CRISP DM 1.0. Data Mining guide," 2000.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten (2009), "The WEKA Data Mining Software: An Update," SIGKDD Explorations, Volume 11, Issue 1.

[6] R. Capon, Y. Tang, R. Forbes, and P. Clark, "A very high resolution model for local fog forecasting," Cost Action 722 Report, 2008.

[7] T. Bergot et al., "Intercomparison of single-column numerical models for the prediction of radiation fog," Cost Action 722 Report, 2008.

[8] B.W. Golding, "Nimrod: a system for generating automated very short range forecasts," Meteorol. Appl. 5, 1-16, 1998.

[9] I. Gultepe and J. Milbrandt, "Microphysical observations and mesoscale model simulation of a warm fog case during FRAM project," Pure Appl. Geophys., vol. 164, 6/7, pp. 1161–1178, 2007.

[10] R. Capon, "Fog forecasting at very high resolution with the Met Office Unified Model," Met Office Forecasting Research Technical Report 444, JCMM Report 149 (available at http://www.metoffice.gov.uk), 2004.

[11] Pasini, V. Pelino, and S. Potestà, "A neural network model for visibility nowcasting from surface observations: results and sensitivity to physical input variables," J. Geophys. Res.106, 14951–14959, 2001.

[12] W. Jacobs and V. Nietosvaar, Foreword. Cost Action 722 Final Report, 2008.

[13] M.F. Al-Roby and A.M. El-Halees, "Data Mining Techniques for Wind Speed Analysis," Journal of Computer Eng., Vol. 2, No. 1, 2011.

[14] G. Li and J. Shi, "On comparing three artificial networks for wind speed forecasting," Applied Energy, vol. 87, no. 7, pp. 2313-2320, Jul. 2010.

[15] C.T. Dhanya and D.N. Kumar, "Data Mining for Evolving Fuzzy Association Rules for Predicting Monsoon Rainfall of India," Journal of Intelligent Systems, Vol. 18, No. 3, 2009.

[16] S. Dong-Jun and J.P. Breidenbach, "Real-Time correction of Spatially Nonuniform Bias in Radar Rainfall Data Using Rain Gauge Measurements," Hydrometeorology, Vol. 3, no. 2, pp. 93-111, 2002.

[17] L. Hluchy et al, "Prediction of significant meteorological phenomena using advanced data Mining and integration methods," Fuzzy Systems and Knowledge Discovery (FSKD), vol. 6. pp. 2998-3002, 10-12 Aug. 2010.

[18] S.N. Kohail and A.M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis," Int. Journal of Information and Communication Technology Res., Vol. 1, No. 3, 2011.

[19] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, and K. Menagias, "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperatures Values," International Journal of Mathematical, Physical and Engineering Sciences, pp. 16-20, 2007.

[20] G. Zazzaro, "An index for local fog forecast by applying Data Mining techniques," Fog Remote Sensing and Modeling (FRAM) Workshop, Dalhousie University, Halifax, Nova Scotia, 21-22 May, 2008.

[21] G. Zazzaro, P. Mercogliano, and F.M. Pisano, "Data Mining to Classify Fog Events by applying Cost-Sensitive Classifier," CISIS 2010, The Fourth International Conference on Complex, Intelligent and SW Intensive Systems, Krakow, Poland, 15-18 February 2010.

[22] G.T. Weymouth, "Dealing with uncertainty in fog forecasting for major airports in Australia," In 4thConference on Fog, Fog Collection and Dew, La Serena, Chile, pp. 73-76, 2007.

[23] F.F. Ebecken, "Fog Formation Prediction in Coastal Regions Using Data Mining Techniques," in International Conf. On Environmental Coastal Regions, Cancun, Mexico, vol. 2, pp. 165-174, 1998.

[24] G. Zazzaro, P. Mercogliano, G. Romano, V. Rillo, and S. Kauczok, "Short Range Fog Forecasting by applying Data Mining Techniques," 2nd IEEE International Workshop on Metrology for Aerospace, at Benevento, Italy, June 3-5 2015, vol. 1, pp. 469-474.

[25] ECMWF. Mars User Guide. User Support. Operations Dep. 2013.

[26] World Meteorological Organization, 2011. Manual on Codes. WMO-No. 306. Volume I.2.

[27] W.T. Roach, "Back to basics: Fog: Part 1—Definitions and basic physics," Weather 49.12, pp. 411-415, 1994.

[28] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2001.

[29] I.H. Witten and E. Frank, "Data Mining. Practical Machine Learning Tools and Techniques," Morgan Kaufmann, 2005.

[30] L.I. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms," John Wiley and Sons, Inc., 2004.

[31] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20(3), pp. 226-239, 1998.

[32] R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, 1993, San Mateo, CA.

[33] B.W.Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta, vol. 405, pp. 442-451, 1975.

[34] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," Technical Report SIE-07-001, 2007.

[35] C. Sammut, G.I. Webb, "Encyclopedia of Machine Learning and Data Mining. II Ed.," Springer, 2017.