BayesNet and Artificial Neural Network for Nowcasting Rare Fog Events

Two different Models for 1-Hour Fog Prediction at Linate Airport

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

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

***Abstract*— Fog represent high impact atmospherical phenomena especially for aviation. In particular, in 2001 the Linate Airport in Milan was interested 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 forecast 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 and other relevant meteorological parameters collected in the Synop message by applying BayesNet and Neural Network algorithms. The performances evaluation shows the complete model for fog events forecasting presents 90% of instances correctly predicted. The work 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; Knowledge Discovery in Meteorological Database Process; Weka; CRISP-DM.

I. INTRODUCTION

Forecasting of adverse weather condition, having high impact on the different phases of the flight (e.g., taxing, landing, take off), is an important issue for air traffic safety. For this reason, many efforts are spent by aviation research for improving the capability to forecast them on different time range. For example, adverse 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, 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 [1]. For these reasons, a fast forecasting is crucial to 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, we introduce a method for fog nowcasting (short-range forecasting of 1 hour) on Linate Airport in Milan using Data Mining techniques. Indeed Data Mining (DM) [2] – 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, is been conducted according to the standard process conceived from the Cross-Industry Standard Process for DM (CRISP-DM) [3]. Every step of the process has been supported by the validation of domain experts. In this work we used the Weka tool (Version 3.6.14) (Waikato Environment for Knowledge Analysis) [4] to carry on DM analysis. In particular, we used the Weka Explorer interface to mine data by applying Bayesian and Artificial Neural Networks 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

 ISBN: 978-1-61208-558-6

model testing. Finally, in Section VI, we present the Evaluation phase and, in Section VII, 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) [5], because the vertical and horizontal resolutions are larger than the corresponding fog scales [6] 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 [1] [7] [8] are unable to treat complex three-dimensional flows due to their poor representation of horizontal heterogeneities [6].

In order to overcome such limitations, dedicated NWP models have been implemented [9] 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 [5]. For this reason, they can be applied only on small domains and on high speed computer [5].

Finally, the statistical methods [10] 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 [11].

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 [12] [13], rainfall [14] [15], cloud [16] and temperature [17] [18].

DM description task is carried on in [19] and [20] 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 [21] 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 [22]. In [23] 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 [24] containing the surface Synoptic observations (SYNOP) provided by Linate meteorological station.

TABLE I. LIST OF METEOROLOGICAL VARIABLES

#	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

 ISBN: 978-1-61208-558-6

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 can be seen. Visibility values below 1 km 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

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

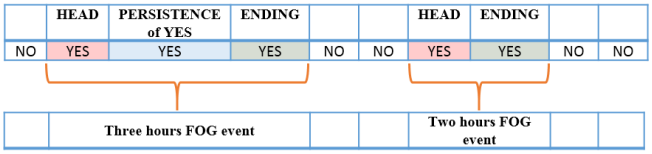

Figure 1. Sequences of recordings.

In Figure 1, 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).

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 [4] [28] 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 2 reports the number of instances of Dewpoint variable in the dataset considered. Dewpoint histogram presents a distribution similar to a Gaussian one.

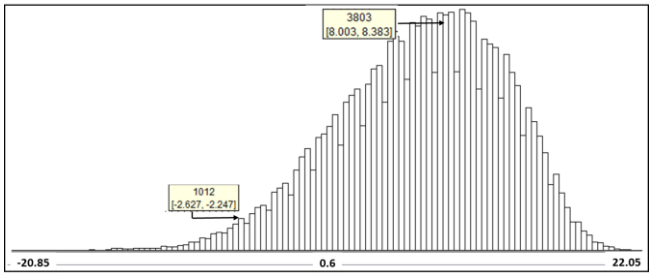

Figure 2. Histogram of instances by Dewpint attribute.

In TABLE II some basic statistics 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.

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.

A. Variables Transformation and Target Class Creation

The meteorological parameters coded according to the World Meteorological Organization (WMO) code tables [25] 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 [26].

The histogram of target class of Figure 3 shows how fog is a quite rare meteorological event on Linate airport: fog occurs about once every 53 events. Target class is unbalanced.

 ISBN: 978-1-61208-558-6

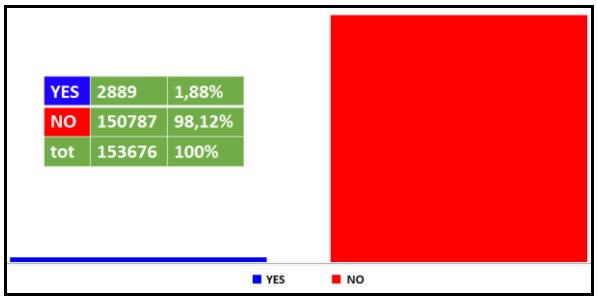

Figure 3. Histogram of instances by class target attribute

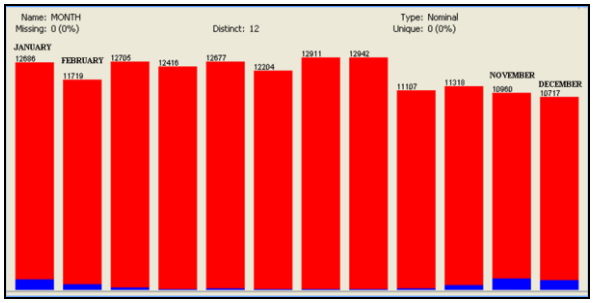

Figure 4. Histogram of instances by Month attribute, from Jan. to Dec.

In order to visualize the distribution of FOG according to the variation of variables, the graph of Figure 4 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.

B. Model Design

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*:

1. *A-model* is used in order to predict the persistence of NO and the discontinuities from NO to YES (heads of new fog events).
2. *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 t0) is greater than 1000 m and *B-model* in the other cases. A summary of this criterion is reported in TABLE III.

TABLE III. RULE FOR MODELS APPLICATION

if visibility at time t0 > 1000m then *A-model* else *B-model*

C. Hold-Out Method and Forecast Sets Preparation

For DM goal, we adopted the working strategy named hold-out method [27]. In this method, the original data with labeled examples is partitioned into two disjoint sets, called the 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 1YEAR, it has 9314 full weather observations and it has roughly the same target class distribution of whole dataset.

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 1YEAR+1.

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

*A*_FOG+1, *B*_FOG+1 aim to train *A-model* and *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 summarizes all of the steps of the Data Preparation phase.

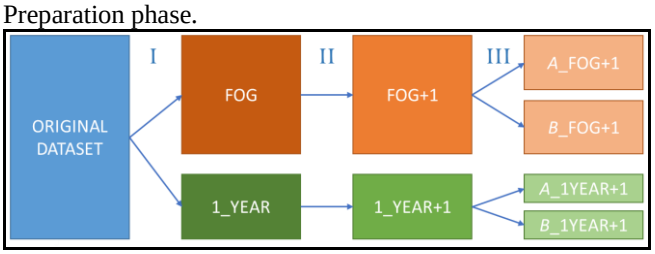

Figure 5. Forecast Sets Preparation Schema

In particular, after the step I, detailed in Figure 5, 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 IV and TABLE V:

TABLE IV. *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 V. *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 3), 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.

 ISBN: 978-1-61208-558-6

Finally, the Data Preparation produces the four datasets presented in TABLE VI, including their sizes.

TABLE VI. 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

V. MODELING

After the Data Preparation follows the Modeling phase, in which the two forecast models are trained and tested.

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.

All obtained prediction 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). The testing of the two 1-hour classification models show good performances, as in next Sections reported.

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.

A. 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 WEKA filter SpreadSubsample that under samples the dataset in order to obtain the same number of FOG+1=”YES” and FOG+1=”NO” instances. This balancing technique is used in order to overcome the class imbalance problem.

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 by applying cross-validation method with k = folds = 10. *A-model* performs on 10-fold cross-validation and it shows the performances included in TABLE VII, TABLE VII, and TABLE IX:

TABLE VII. *A-MODEL* EVALUATION

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

TABLE VIII. *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 IX. CONFUSION MATRIX OF *A-MODEL*

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

A-model shows the performances on *A*_1YEAR+1 Test Set included in TABLE X, TABLE XI, and TABLE XII.

TABLE X. *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 XI. *A-MODEL* 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 XII. CONFUSION MATRIX OF *A-MODEL* ON *A_1YEAR+1*

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

In this test, we analyze 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”).

B. The B-model

The B-classifier is an Artificial Neural Network (ANN) 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”.

The *B-model* is trained by using the MultilayerPerceptron algorithm of WEKA, 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 XIII, TABLE XIV, and in TABLE XV:

TABLE XIII. THE B-MODEL EVALUATION

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

TABLE XIV. THE *B-MODEL* 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

 ISBN: 978-1-61208-558-6

TABLE XV. 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 XVI, TABLE XVII, and TABLE XVIII.

TABLE XVI. 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 XVII. THE *B-MODEL* 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 XVIII. CONFUSION MATRIX OF THE *B-MODEL* ON *B_1YEAR+1*

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

In this test we analyze 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). The TABLE XIX shows the distribution of such False Positives by Month attribute. 74% of False Positive instances occur in [September, January].

TABLE XIX. 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		

The Figure 6 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).

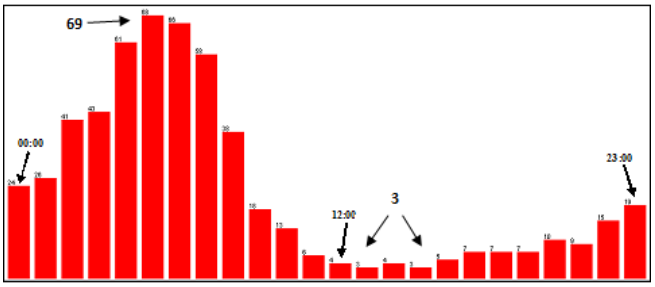

Figure 6. Histogram of false positives by Hour attribute

In addition, about 75% of False Positive instances occur when visibility ranges from [1200m, 4500m] (range of Visibility attribute). About 70% of False Positives occur when the 'Height of base of cloud' attribute is in [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. The TABLE XX shows the distribution of False Positives by 'Present Weather' attribute.

TABLE XX. 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	

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.

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.

Finally, 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. CONCLUSIONS

This paper reports the description of a statistical tool to forecast in a very rapid time the occurrence of low visibility

 ISBN: 978-1-61208-558-6

events over the airport area. This method is essentially based on the use of an historical time series of SYNOP data available over Linate airport and on the DM techniques. SYNOP are a meteorological data message available in many airport, 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 included 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. *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 *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, 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] 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.

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

[3] P. Chapman et al., "CRISP DM 1.0. Data mining guide", 2000.

[4] 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.

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

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

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

[8] I. Gultepe and J. Milbrandt, "Microphysical observations and mesoscale model simulation of a warm fog case during FRAM project", Pure Appl. Geophys. 164, 7/8, this issue, 2007.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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, No3, 2011.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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, Volume: pp 460-465.

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

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

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

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

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

 ISBN: 978-1-61208-558-6