

A DSS for Assessing the Impact of Environmental Quality on Emergency Hospital Admissions

Daniele Toscani¹, Francesco Archetti^{1,2}

¹Consorzio Milano Ricerche
Milan, Italy
<surname>@milanoricerche.it

Luigi Quarenghi², Federica Bargna², Enza Messina²

²DISCo – Department of Computer Science, Systems and
Communication
University of Milan Bicocca
Milan, Italy
<surname>@disco.unimib.it

Abstract— In this paper we present a Decision Support System (DSS) aimed at forecasting high demand of admission on health care structures due to environmental pollution. The computational engine of the DSS is based on Autoregressive Hidden Markov Models (AHMM). We estimate the forecasting model by analyzing the historical daily average concentrations of pollutants and the number of hospital admissions, collected from multiple data sources. Given the actual concentration of different pollutants, measured by a network of sensors, the DSS allows to forecast the demand of hospital admissions for acute diseases in the following 1 to 6 days. We tested our system on cardiovascular and respiratory diseases in the area of Milan. The performances of our system, compared with multiple linear regression, show that AHMM are a robust approach to capture the connections among health and environmental indicators.

Keywords- Time series analysis, Health-environment correlation, Hospital demand planning

I. INTRODUCTION

The evaluation of the quality of the environment is important for health impact assessment, which in turn is crucial for taking actions for protecting the public health. Many epidemiological studies have shown that acute or chronic health effects can be associated with high concentrations of environment pollutants. This paper stems from the research project LENVIS (*Localised ENVironmental and health Information Services for all*) aimed at designing a Decision Support System (DSS) based on statistical learning. The main aim of the project is to leverage on-line environmental monitoring into alerts during episodic pollution events, in order to inform people of the current environmental situation, give warnings to health care providers about peaks in requests of hospitalization, support public authorities in deciding which actions have to be carried out (e.g. prevent risks by reducing excess pollution and minimizing exposure, prepare hospitals to peaks of emergency admissions).

The case study considered in this paper is focused on air quality whose data are collected through a network of monitoring stations which measure the daily concentration of pollutants in the urban area of Milan. These data, together with the data regarding hospital admissions in the same area, are used to build a forecasting model for predicting the number of hospital admission requests in the next 1 to 6 days. For this

purpose we considered a particular type of state space model, namely Autoregressive Hidden Markov Model (AHMM). Sect II gives an overview of the current state of the art in environmental and health studies. Sect. III outlines the software and functional architecture of our DSS. Sect. IV is devoted to describe our case study: air pollution in the town of Milan. Sect. V presents experiments result, comparing the accuracy of AHMM with the traditional Multiple Linear Regression. Sect. VI discusses conclusions and future work.

II. STATE OF THE ART

The WHO quality guidelines [1] assert that health effects from exposure to air pollutants may vary from increased premature deaths [2] to aggravation of respiratory and cardiovascular illness, as decreased lung function and symptomatic effects (e.g. acute bronchitis), new cases of chronic bronchitis and heart attacks. These effects are particularly observed in children, elderly [3] and individuals with heart or lung conditions, leading to hospitalizations and access to emergency services. In this paper we focus on the typical health effects of air pollution: respiratory diseases (asthma) and cardio-vascular diseases (myocardial infarction, ischemic cardiopathy, deep vein thrombosis). The connections among pollution and health have already been addressed in a number of epidemiological studies, mainly focused on statistical analysis [4][5][6]. We use the findings of these studies to identify the most relevant pathologies, moreover, i.e. to correlate peaks of hospitalization and pollution levels, is different and consequently our methods are different from those of epidemiological studies. Notably, we do not need to use dose-response relationships. The key components of our approach is time series analysis, in order to understand how hospitalization levels change in a specific area as air pollution levels fluctuate. The primary advantage of time-series studies is that many factors that could potentially confound the association between air pollution and health are constant during the study period and are therefore not relevant. The only factors likely to vary with daily mortality and morbidity are environmental and meteorological conditions, which typically are accounted for in the statistical analysis.

A number of techniques have been developed for the prediction of pollutant concentrations or health indicators [7]. Essentially, forecasting approaches can be grouped in empirical

This research has been partially supported by the project LENVIS - *Localised ENVironmental and health Information Services for all*, European Community's 7th Framework Programme (FP7/2007-2013) under grant agreement n° 223925.

models, fuzzy logic based systems, data driven statistical models and model-driven statistical learning methods. We focus on this last category, to which belong state space models and Bayesian networks. In particular, we analyze historical data to infer a mathematical model that links variables and can be used in subsequent analysis. The “learning from data” and the probabilistic links among quantities make model driven methods able to exploit patterns and relations that cannot be explicitly handled by rule-based or fuzzy logic systems. Furthermore, in contrast to the black box approaches such as neural networks, the model-driven approach explicitly deals with observable variables and can be adapted to increase accuracy and answer complex and inter-temporal queries. Consequently, model-driven methods can significantly outperform data-driven approaches.

III. SYSTEM ARCHITECTURE

The objective of our DSS is to forecast demands of hospital admissions given the current pollution and to deliver information to e.g. public authorities and hospital managers. The DSS operates in 3 phases, as shown in Figure 1. :

- Access to the environmental and health data through the data integration infrastructure;
- Training of the forecasting model (AHMM) through analysis of historical data;
- Forecast of the future value for health indicator (number of hospitalizations) given the actual environmental conditions.

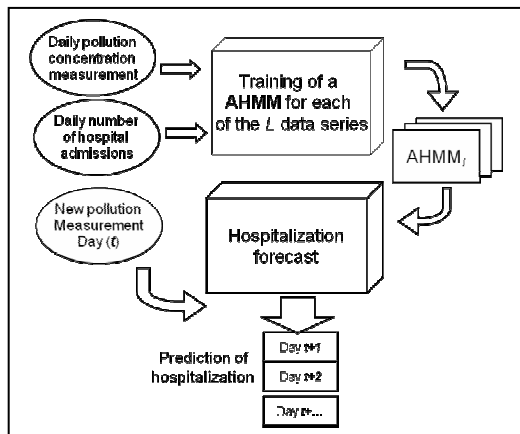


Figure 1. Workflow of our DSS

Figure 2. depicts the two main components of the system: the *Data Integration Infrastructure* (DII) and the *Health Impact Decision Support System* (HIDSS), together with the main data flows from data providers to the system. The system is implemented in Java and part of the computational engine in Matlab. External applications and user interfaces control the DSS and collect the outputs through web services and API (Application Programming Interfaces).

A. Infrastructure to support data access and integration

Input data are collected through web services and local databases. The DII is a software component that provides

integrated access to multiple, heterogeneous and distributed data sources. Its functionalities are: object-oriented representation of the types of data supported; submission of structured queries; return of the query result as a searchable and navigable data structure. It allows executing cross-source queries on temporal, spatial and logical intervals, supporting data analysis and presentation activities. Each query, despite the traditional SQL syntax, specifies a target but not the data sources from which to extract the data. The platform defines the data types that the data sources can provide. Each data type is a structured object, containing multiple fields (e.g. for a sensor reading, the source of the data, the value collected, the timestamp to which it refers etc.) and it is linked to other data types through hierarchical relations. Each query also specifies the constraints on results (temporal intervals, values of some attributes...). This querying mechanism hides the heterogeneity and distribution of the data sources. Moreover, it is responsibility of the DII to identify, select and query all the sources needed to reply to user requests and also to prepare and present the output in the common format.

The DII is fully extensible both in the type of sources and integration mechanisms, giving access to relational databases, data warehouses, structured and semi-structured data, web services etc. A fundamental feature is the possibility to reply to queries not only accessing to persistent data but also using streaming data produced online by sensors (e.g. sensor networks) or by the models of the HIDSS described in Sect. III.b, which perform online inference (forecast).

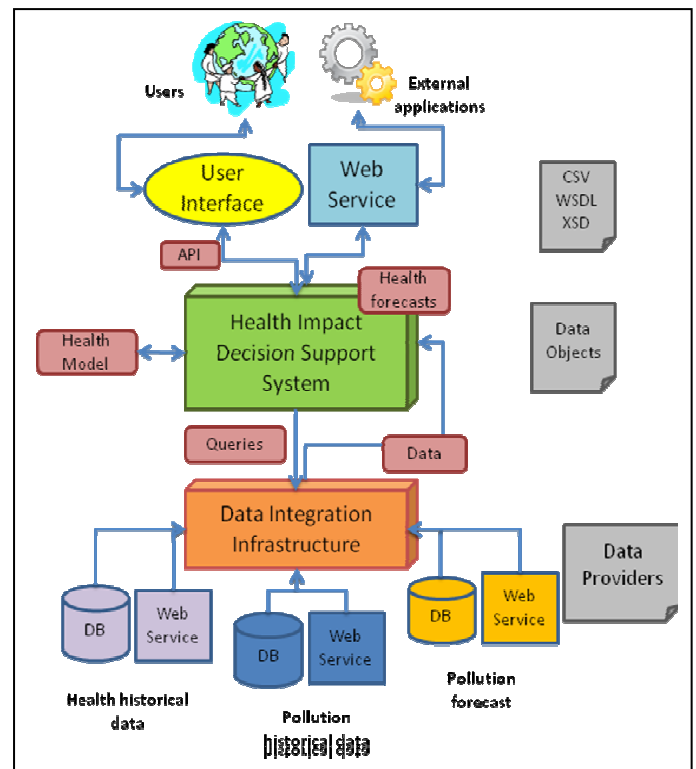


Figure 2. Main components and data flows

We collect through the DII the environmental data and health indicators (Sect. IV.a and IV.b). For each series, we obtain a scale-free summary of changes in the data during time by calculating the *log-variation* among two consecutive values v_{t-1} and v_t , i.e. $\log(v_t/v_{t-1})$. Log-variations are the observations o_t for our AHMM (Sect III.b).

B. The Computational Engine: Hidden Markov Models

The core of our DSS is the computational engine. While the literature shows mostly the application of regression models, the authors feel that the dynamic of the target process is better captured by probabilistic state-space models, namely Hidden Markov Models (HMM) [8]. In fact, while statistical analysis of time series, like multiple linear regression, is based on local information and assumes stationarity, or weak stationarity, of the data, health data shows asymmetries between growth and decline periods and it is often difficult to distinguish between long-term trends and noise. As result, a forecasting model is valid only for a short period. It is then important to employ models able to capture different trends and dynamically switch among them while the environmental conditions change. Instead of considering a single global model, state space models are concerned which include adjusting several local models for the different time series regimes.

When data are generated by a process that we cannot directly observe, or is too complex, HMM allows ignoring or partially simplify its nature, focusing the attention on the data generation process, which is indeed our final objective. A HMM is a stochastic process whose evolution is governed by a Markov Chain whose state variable can assume a finite number of values $s_j, j=1,2,\dots,N$, which are often called *states* and cannot be directly observed. Each s_j corresponds to one of the possible “conditions” in which the process can be. The model is characterized by a probability distribution for the initial value of the state $\pi(s_j)$, a set of state transition probabilities A whose elements $a_{ij}=P(s_j/s_i)$ measure the probability of being in state s_i at time t and s_j at $t+1$. We apply *Autoregressive HMM* (AHMM), a class of HMM which assumes that an observation depends on the current state and also on the previous p values. The formulation of the model is completed by a set B of Probability Density Functions (PDF) which describe the generation (*emission*) of the observed values in each state. The observation o_t at time t is generated by s_j according to:

$$o_t = b_j(O) = P(o_t | s_j, o_{t-1}, \dots, o_{t-p}) \quad (1)$$

K subsequent observations are grouped in vectors, which define an autoregression process of order p :

$$o_k = \sum_{i=1}^p r_i o_{k-i} + e_k, \quad k = 1, 2, \dots, K \quad (2)$$

Where e_k is a Gaussian noise component, with zero mean and variance σ^2 , and r_i ($i=1,2,\dots,p$) are the autoregression coefficients. Note that, the PDF of the observed variable O for state s_j is a mixture of M components:

$$b_j(O) = \sum_{m=1}^M c_{jm} b_{jm}(O), \quad \sum_{m=1}^M c_{jm} = 1, \quad c_{jm} \geq 0 \quad (3)$$

Where c_{jm} are a set of weight, whose sum is 1, and $b_{jm}(O)$ is the observation density for state j and m^{th} mixture component, computed as follows:

$$b_{jm}(O) = \left(\frac{2\pi}{K}\right)^{-K/2} \exp\left(-\frac{K}{2} \delta(O, r_{jm})\right) \quad (4)$$

Where

$$\delta(O, r_{jm}) = R_{r_{jm}}(0)R(0) + 2\sum_{i=1}^p R_{r_{jm}}(i)R(i) \quad (5)$$

$$R_{r_{jm}}(i) = \sum_{n=0}^{p-1} r_{jmn} r_{jmn+i-1} \quad (r_{jmn=0}=1), \quad 1 \leq i \leq p \quad (6)$$

$$R(i) = \sum_{n=0}^{K-i-1} o_n o_{n+i}, \quad 0 \leq i \leq p \quad (7)$$

r_{jmn} is the autoregression coefficient for state j , mixture component m and element n of the autoregression process. $R_{jmn}(i)$ is the autocorrelation of the autoregression coefficients for state j and mixture component m and $R(i)$ is the autocorrelation of the observations.

The model is fully characterized by the set of parameters $\lambda = [\pi(s_j), a_{ij}, r_{jmn}, R(i)]$, which have been estimated using the customized version of the Baum-Welch algorithm described in [9]. Since there is no analytical process to estimate the optimal order p of the autoregression process, given a sequence of T training values v_1, \dots, v_T , we estimate multiple autoregression processes θ_p for different p by maximizing the conditional distribution $f(v_1, \dots, v_T | \theta_p)$, i.e. their *likelihood* (MLE - *Maximum Likelihood Estimation*). Then, we select the optimal p through the *AIC* (*Akaike Information Criterion*) [10]. AIC is the selection criterion applied commonly for its simplicity and good performance [11]; it favours models with high likelihood while penalizing complex model structures. The initialization of autoregression parameters is performed by analyzing the relationship between autoregression parameters and the autocorrelation function of the observed data by solving a system of Yule-Walker equations [11].

There is a strong relation between HMM and *Dynamic Bayesian Networks* (DBN), i.e. probabilistic models that allow representing temporal dependencies among random variables [12]. The DBN in Figure 3. graphically represents the variables of an AHMM for T observations: the basic structure is repeated for every time step; the state variable influences the value observed at the same time, according to the observation model B , and the state at time $t+1$ according to transition model A . Figure 3. shows the feature of AHMM in which, unlike simple HMM, observations o_t and o_{t+1} are linked by the autoregression process.

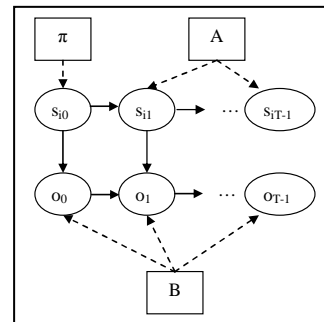


Figure 3. Graphical structure of the DBN representing an AHMM

The formulation of AHMM described here can deal only with uni-variate series. To manipulate multiple pollutants and health indicators at the same time we use multiple models, one for each series, as in [9]. We capture the relations among series by estimating the Cholesky decomposition R_D of their correlation matrix [12].

The output of learning is the “Health Model” (HM), which encapsulates the “probabilistic relations” that will be applied to forecast health data. In the forecast phase, using as input the actual concentration of pollutants measured by sensors, the Health Model is applied to forecast health indicators for the subsequent days. The generation of data is as follows: each AHMM is applied to sample T data. For each time $t=1,2,\dots,T$ we produce a vector of forecast (one value for each series): given the current state s_t we sample the next state s_{t+1} according to the state transition probability and the observation o_t from the distribution $b_t(O)$. Finally, as in [14], we restore the correlations among series by multiplying the forecasts at the same time by R_D .

IV. CASE STUDY: THE MILAN AREA

We selected as case study the air pollution in the metropolitan area of Milan, which is the Italian area with the highest population (10th in Europe), very high population density and it is affected by air stagnation; all these conditions have brought to considerable air pollution problems. This represents the ideal test bed for our DSS, since we can observe a high number of pollution peaks.

A. Environmental Data

We consider as environmental data the daily average concentration of air pollutants. In the city of Milan are installed 9 monitoring stations whose position is shown by the circles in the map of Figure 4. Each station is equipped with a variable number of sensors, for a total of 37 sensors in the whole town. Each sensor measures the concentration (in $\mu\text{g}/\text{m}^3$) of one among: benzene, nitrogen dioxide, sulphur dioxide, carbon monoxide, nitrogen oxide, total nitrogen oxide, ozone, PM10 (Particulate Matter), PM2.5, TSP (Total Suspended Particulate). The station calculates each hour the average concentration and sends it to a control centre, where the data are manually validated to filter outliers and aggregated to obtain a daily measure. These data are publicly available at the web site of the environment protection agency of the Lombardy region (ARPA Lombardia). http://www.arpalombardia.it/qaria/doc_RichiestaDati.asp. Most of the data are available from the 80s-90s, but in many stations some data are available only after year 2000, in one case from 2007. The series of historical data are not complete, since some stations have been under maintenance for months or have been switched off. In building the DSS, we had to take into consideration this feature, since not all the data series could be analyzed together for a given time period.

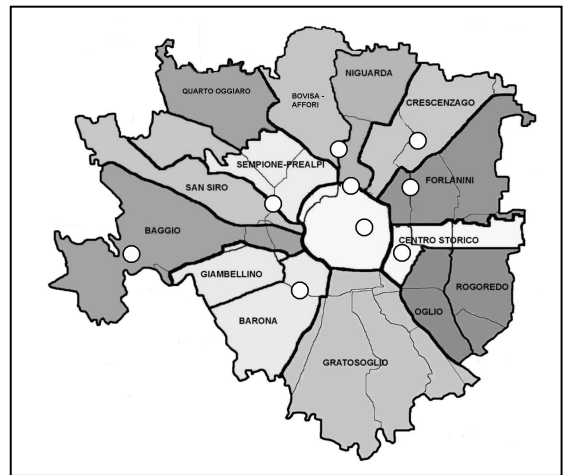


Figure 4. City map of Milan. Circles show air pollution monitoring stations

B. Health Data

The health indicator that we address in our analysis is the daily number of hospitalizations in the town of Milan for diseases whose acute occurrence can be related to pollution. In particular, two classes of disease have been considered: 1) respiratory (asthma) and 2) cardio vascular (myocardial infarction, ischemic cardiopathy and deep vein thrombosis). These data are collected by the local government of the Lombardy region and published on <http://www.aleao.it/>; from this source we extracted two data series, one for each class of diseases, recording for each day of the period January 1998 – December 2007 the number of persons that have been admitted to hospitals in Milan, with a diagnosis in one of the two classes.

C. Experimental Setup

We grouped pollutants into three classes of homogeneous substances; for each class we selected a representative series, resulting in 9 triples of series whose mutual correlation coefficient is <0.8 . Each triple has been used to predict cardiovascular admissions and, separately, respiratory ones. The resulting configurations of the experiments (E1..E18) are detailed in TABLE I.

We analyzed series of about 3000 values. We determined experimentally the optimal structure of the AHMM, which achieves highest model likelihood with respect to training data, having 2 states, 5 mixture components and an autoregression process with order variable from 2 to 5. In each experiment we trained the model on a sliding window of 500 observations and we forecast the number of hospitalizations from 1 to 6 days. For each series we measure the performance by calculating the *Mean Absolute Percentage Error* (MAPE) between the real number of hospitalizations and the forecast produced by AHMM. In order to assess the model reliability, the forecast is repeated 6 times for each model. We present the results in terms of global MAPE, calculated as mean of all the test executed with the same series. For evaluating the performance of our forecasting model we used as benchmark the *Multiple Linear Regression* (MLR) [15], a multivariate statistical technique widely used to capture the linear correlations

between some predictor variables $v_{j..v_{L-1}}$ (i.e. concentrations of pollutants) and a single dependent variable v_L (response variable, i.e. health indicator).

TABLE I. CONFIGURATION OF DATA SERIES IN EACH EXPERIMENT

Pollutants	Experiment / Pathology
Nitrogen oxides, Sulfur dioxide, PM 10	E1 Card., E2 Resp.
Nitrogen oxides, Ozone, PM10	E3 Card., E4 Resp.
Nitrogen oxides, Carbon monoxide, PM 10	E5 Card., E6 Resp.
Nitrogen oxides, Benzene, PM10	E7 Card., E8 Resp.
Nitrogen dioxide, Sulfur dioxide, Benzene	E9 Card., E10 Resp.
Nitrogen dioxide, Sulfur dioxide, Total PM	E11 Card., E12 Resp.
Nitrogen monoxide, Ozone, Total PM	E13 Card., E14 Resp.
Sulfur dioxide, Ozone, Carbon monoxide	E15 Card., E16 Resp.
Ozone, Total PM, PM10	E17 Card., E18 Resp.

V. EXPERIMENTAL RESULTS

The wide set of computational experiments performed has shown that AHMM achieve low average error on forecast than MLR. In the DSS, these statistics can help the user in selecting the “best” mix of pollutants, i.e. the set of sensors whose monitoring can give the low error in forecast and hence the best advice on health indicators. Figure 5. and Figure 6. show the results on cardio-vascular diseases; experiments for AHMM are labelled with “H”, for MLR with “R”. Each experiment is constituted by multiple tests, as described in Sect IV.c; for each experiment we report the minimum, maximum and average value of MAPE. AHMM achieves a mean value from 7.424 to 19.64, while the best result for MLR is 15.86. While MLR achieves a relatively high error, which is approximately constant in all the experiments, the result for AHMM vary depending on the experiment. This variation can be explained by the nature of the learning process of AHMM. In fact, while MLR is a deterministic model, which can be trained on a small quantity of data, training procedures for AHMM are non deterministic optimization techniques; the quality of the resulting model depends on the reaching of the local maxima and, definitively, on the nature of data. The performance of AHMM can be improved by executing multiple times the learning process and selecting the maximum likelihood model, or even changing the structure of the model and the quantity of data. Another remarkable feature of AHMM is that in most experiments the maximum and minimum value for MAPE in different tests is very close to the mean; this shows the robustness of the approach, which demonstrates that the low error can be achieved in different conditions.

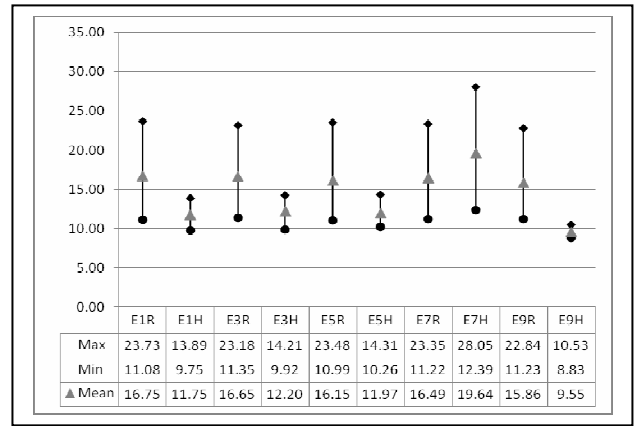


Figure 5. MAPE of experiments E1..E9 on cardio-vascular diseases

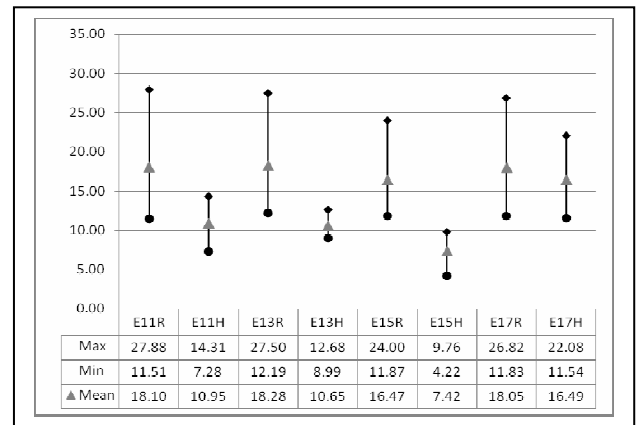


Figure 6. MAPE of experiments E11..E17 on cardio-vascular diseases

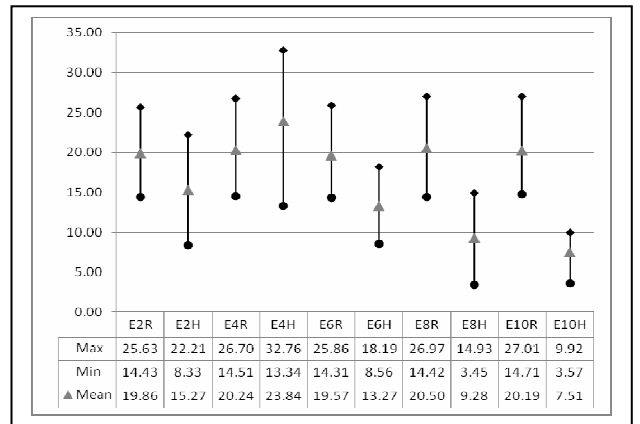


Figure 7. MAPE of experiments E2..E10 on respiratory diseases

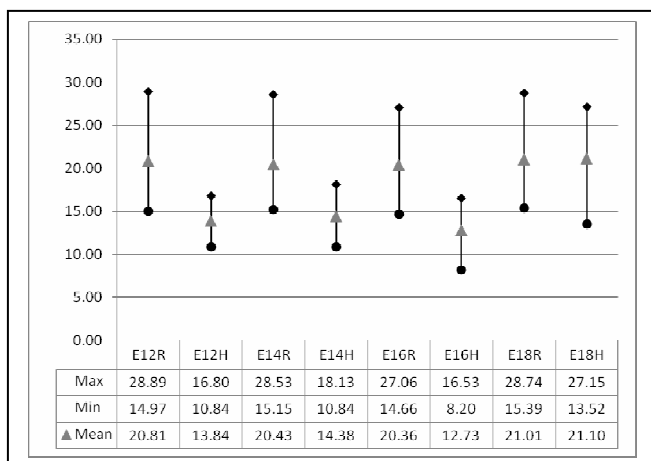


Figure 8. MAPE of experiments E12..E18 on respiratory diseases

Figure 7. and Figure 8. show the MAPE in forecast respiratory diseases. As for cardio-vascular, in most cases the MAPE for AHMM is lower than for MLR, ranging from 7.505 to 23.84 for AHMM and from 19.56 to 21.01 for MLR. The variation among maximum and minimum MAPE in the test for each experiment is generally lower for AHMM, showing as in the previous case that this model can provide more stable results.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we present a DSS for the collection of environmental and health data from multiple sources and their analysis through Autoregressive Hidden Markov Models. In particular, given the measurements of the concentration of some pollutants in a given time period, the main objective is to forecast the number of hospitalizations for acute cases of the related pathologies. The results show that AHMM outperforms traditional statistical methods in the forecast.

The interest in our work has been expressed by professional users represented by researchers of European universities in the area of environment, computer science and health, together with medical staff of public hospitals. The type of environmental and health information that is typically available is related to diseases and epidemiological statistics. This shows that the short term link between environment and health has not been exploited yet. In fact, information is sometimes available on dedicated web sites, but in many case specialized papers and journals are the only source of information. The substantial innovation of this paper is to provide an automatic system that allows to have continuously updated forecasts about the demand for health services, i.e. hospitalizations and emergency accesses..

Our work is still in progress: the links with experts in environment and health are supporting us in understanding the data, choosing the models and interpreting results. Further analysis has to be performed in order to expand the possibilities of data integration and the robustness of the software

components of the computational engine. We are actually expanding the application field of our DSS to water pollution, to monitor pathologies of the digestive system, skin and subcutaneous tissue.

ACKNOWLEDGMENTS

A special thanks goes to Prof. Daniela Mari, of the University of Milan and IRCSS Istituto Auxologico Italiano, which supported our work in the selection and interpretation of medical data, and Dr. Luca Grappiolo, of IRCSS Istituto Auxologico Italiano, which provided us support in the design of data collection procedures and the DII.

REFERENCES

- [1] WHO, "Air quality guidelines - global update 2005", World Health Organization Publications, Bonn, Germany, Tech. Rep., 2005.
- [2] C. Arden Pope III, M. Ezzati and D. W. Dockery, "Fine-particulate air pollution and life expectancy in the United States," *The New England Journal of Medicine*, vol. 360, no. 4, pp. 376-386, Jan., 2009.
- [3] S. Dubowsky Adar, D. R. Gold, B. A. Coull, J. Schwartz, P. H. Stone and H. Suh, "Focused exposures to airborne traffic particles and heart rate variability in the elderly," *Epidemiology*, vol. 18, pp. 95-103, 2007.
- [4] G. A. Wellenius, J. Schwartz and M. A. Mittleman, "Particulate air pollution and hospital admissions for congestive heart failure in seven united states cities," *The American journal of cardiology*, vol. 97, no. 3, pp. 404-408, 2006.
- [5] M. Medina-Ramon, A. Zanobetti and J. Schwartz, "The effect of ozone and PM10 on hospital admissions for pneumonia and chronic obstructive pulmonary disease: a national multicity study," *American Journal of Epidemiology*, vol. 163, no. 6, pp. 579-588, 2006.
- [6] P. Bellini, M. Baccini, A. Biggeri and B. Terracini, "The meta-analysis of the Italian studies on short-term effects of air pollution (MISA): old and new issues on the interpretation of the statistical evidences," *Environmetrics*, vol. 18, no. 3, pp. 219-229, 2007.
- [7] M. Dong, D. Yang, Y. Kuang, D. He, S. Erdal and D. Kenski, "PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9046-9055, 2009.
- [8] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. of the IEEE*, 1989, pp. 257-286.
- [9] E. Messina and D. Toscani, "Hidden markov models for scenario generation," *IMA Journal of Management Mathematics*, vol. 19, pp. 379-401, Oct., 2008.
- [10] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd International Symposium on Information Theory*, 1973, pp. 267-281.
- [11] R. S. Tsay, *Analysis of financial time series*, John Wiley & Sons, 2005.
- [12] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2002.
- [13] N. J. Higham, "Computing the nearest correlation matrix", Manchester Centre for Computational Mathematics, Manchester, England, Tech. Rep. 389, 2000.
- [14] K. Hoyland, M. Kaut and S. W. Wallace, "A Heuristic for Moment-Matching Scenario Generation," *Computational Optimization and Applications*, vol. 24, no. 2-3, pp. 169-185, 2003.
- [15] D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, Wiley, 2006.