Markov-based approaches to support policies makers in environment and healthcare

Antonio Candelieri
Francesco Archetti
Ilaria Giordani
Gaia Arosio

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference

Candelieri, Antonio; Archetti, Francesco; Giordani, Ilaria; and Arosio, Gaia, "Markov-based approaches to support policies makers in environment and healthcare" (2012). International Congress on Environmental Modelling and Software. 302.
https://scholarsarchive.byu.edu/iemssconference/2012/Stream-B/302

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Markov-based approaches to support policies makers in environment and healthcare

Antonio Candelieri a, Francesco Archetti a, b, Ilaria Giordani b, Gaia Arosio a
a University of Milano-Bicocca, Italy - Department of Information, Systems and Communication (DISCo)
antonio.candelieri@unimib.it, gaia.arosio@gmail.com
b Consorzio Milano Ricerche, Italy
archetti@milanoricerche.it, giordani@milanoricerche.it

Abstract: Perceived need of big cities is focused on policies limiting pollution from cars exhaust, energy production, industrial manufacture, with the aim to reduce the impact on both environment and citizens health/living. Decision support is mandatory, aiming at monitoring the city's status, evaluating policies effectiveness and estimating/predicting impacts of environment on health in order to plan more rationale resources allocation and care provision.

We present a multi-period analysis performed through Markov-based approaches and prove how these methods may support decisions in the management of big cities, by integrating environment and health related issues. The study has considered historical pollutants data (2002 to 2011) for assessing air quality in Milan, according to the most updated European Directive; approaches have been used to evaluate the effectiveness of a traffic limitation policy brought into force in the 2007 (the Ecopass).

Although several epidemiological studies have provided evidence of a relationship between air quality and mortality/morbidity due to cardiovascular and respiratory diseases, the prediction of emergency hospital admissions in the short term is yet difficult to achieve, even if crucial for a rational healthcare management. In this paper we propose to use Markov-based techniques in order to obtain a model to forecast hospital admissions, in the short term, according to pollution level.

This work has been performed within the European project Lenvis (Local ENVIronmental Services), a collaborative network of services able to retrieve and analyze heterogeneous and geographically dispersed data sources in order to deliver environment and health information (www.lenvis.eu). The Markov-based models, trained and validated on real data, have been deployed into the Lenvis' Health Impact Decision Support System (HIDSS) and made accessible, as services, to decision makers and users accounted. The system has showed its usefulness both for environment authorities and healthcare stakeholders.

Keywords: Markov models, emergency hospital admissions forecasting, pollution control policies, healthcare management.

1 INTRODUCTION

Many research studies focused on the analysis of the influence of environmental factors on health risk. As recently reviewed by Sun et al. [2010], epidemiologic studies on different populations provided pollutant levels and exposure models for
the estimation of the increased risk of morbidity/mortality depending on the pollutant concentrations. A recent report of the World Health Organization (WHO) [2005] indicates a 0.11% increase in mortality, due to cardiovascular or respiratory diseases, for a 10μg/m³ increase in the daily average of PM10 that is the concentration of particles with a diameter lower than or equal to 10 micrometers in the atmosphere. Fine particles are generally produced by industrial and residential combustion and also vehicle exhaust. Exposure to fine particles mainly affects health of frail or sensitive population segments, such as children, the elderly, individuals with heart or lung, acute or chronic, diseases. The most known effects of the exposure to PM, as reported by Zeger et al. [1999], are:

- increase in premature deaths of “frail” persons,
- worsening of lung function and respiratory and cardiovascular diseases, usually requiring emergency visits and eventual hospitalizations,
- onset of new cases of chronic bronchitis and heart attack events,
- modifications to the lung structure and natural defence mechanisms.

European directives, i.e., the European Directive on ambient air quality and cleaner air for Europe (DIRECTIVE 2008/50/EC, [2008]), mandate structural modifications for sustained improvements and encourage strategies for supporting both short term decisions (e.g., temporary traffic limitations) and warning/alerting systems toward specific frail segments of population. Respect to this, perceived need of big cities is focused on policies limiting pollution from cars exhaust, energy production, industrial manufacture, with the aim to reduce the impact on environment and citizens health/living. Decision support is mandatory in order to:

- continuously monitor the city’s status,
- evaluate the effectiveness of pollution control/limitation policies,
- estimate/predict impacts of environment on health in order to plan more rationale resources allocation and care provision strategies.

While epidemiological studies proved to be useful to support a posteriori evaluation over an extended period of time, the informational need related to monitor and forecast the environmental impact on health (e.g., in terms of emergency hospital admissions), in particular in the short term, is not yet fulfilled, even if crucial for a more rational healthcare management. However, some indications are given in literature; Zanobetti et al. [2009] report a two-day averaged increase equal to 10μg/m³ in PM2.5 level is associated to an increase of 1.89%, 2.25%, 1.85% and 2.07% in cardiovascular disease, myocardial infarction, congestive heart failure and respiratory admissions, respectively, in United States. More recently, Blangiardo et al. [2011] pointed out that a 10μg/m³ daily increase in PM10 concentration produces a 0.9%, 0.8% and 1.3% increase in hospital admissions, in London, due to cardiac diseases, myocardial infarction and respiratory disease, respectively.

The European FP7 project Lenvis (Localized ENVIronmental Services, website: www.lenvis.eu) aims at designing and developing an innovative and collaborative decision support network that integrates location-based data and services in order to support stakeholders in exchanging environmental and health related information in a European wide mesh of collaborating services. Lenvis supports information and knowledge sharing among local authorities, citizens and environmental and healthcare stakeholders, moreover it faces to overcome the limits of the traditional epidemiological studies by providing services for monitoring and forecasting environment and health related issues also in the very short term (e.g., emergency hospital admissions in 1-5 days). The Health Impact Decision Support System (HIDSS) is the Lenvis’ component devoted to mine environmental and health data to discover patterns to use for risk assessment and forecast.

In this paper we present the most relevant results obtained by the application of Markov-based strategies for designing and developing computational services to support decision making of environment and health related stakeholders. With respect to the state of the art, some Bayesian approaches have been already proposed, (i) assessing the effect of “episodes” (sustained periods of high pollution) on the risk of cardio-respiratory admission in London by comparing 2003
data with 2002, as shown by Blangiardo et al. [2011] and (ii) pointing out that higher impacts of pollution level on hospital admissions, by taking into account several national datasets from United States, is in winter, as reported by Bell et al. [2009].

We propose Markov-based strategies to:

- model modifications in pollution level during winter seasons (2002 to 2010) in Milan and use the resulting model to evaluate the effectiveness of a traffic limitation policy brought into force in 2007;
- link pollution levels and emergency hospital admissions to forecast the trend of the last ones in order to support a more rational resources planning.

These solutions implements services of HIDSS that provides the following facilities:

- supporting environment related authorities to design and evaluate emissions limitation policies,
- allowing hospital managers to plan resources in advance according to admissions trend forecast in the short term,
- assisting General Practitioners to predict adverse events and prevent hospitalization, reducing costs while improving patient quality of life.

The rest of the paper is structured as follows: in section 2 the available datasets and the pre-processing activities are described, both for environmental and health data; in section 3 we basics of Markov-based inference are summarized and our approach is presented; in section 4 results related to the evaluation of the effectiveness of the Ecopass control policy and the performances of emergency hospital admissions forecast are summarized; finally we report some conclusions.

2 DATASET AND PREPROCESSING

The study is based on two different datasets of environment and health related data, respectively. Both these two datasets refer to the city of Milan but have a relevant difference in terms of data acquisition and time horizons. In particular, the pollution levels are acquired through a sensors network deployed in Milan by the local authority for the environment protection (Agenzia Regionale per la Protezione dell’Ambiente, ARPA Lombardia) while the emergency hospital admissions have been gathered from 2 health centres and the emergency ambulance service "118" by directly involving clinical stakeholders. Details about the two datasets and the necessary data preprocessing steps are reported in the following.

2.1 Pollution level data

The local environment protection agency (ARPA) has installed nine automatic stations in Milan to continuously monitor the following pollutants: Benzene (Be), Carbon monoxide (CO), Nitrogen oxide (NO), Nitrogen dioxide (NO2), Total nitrogen oxide (NOx), Ozone (O3), Particulate Matter 10 and 2.5 (PM_{10} and PM_{2.5}), Total suspended particulate (TSP), Sulphur dioxide (SO2).

The period from 1st January 2002 to 30th April 2011 has been considered as time window for the pollution-related analysis presented in this paper. The ARPA data sources have been integrated in Lenvis, therefore all the environment related data, both historical and updated, are available to the Lenvis accounted users through an user-friendly interface (gadget) of the web portal (www.portal.lenvis.eu).

2.2 Emergency hospital admissions data

Health data is not automatically acquired as for pollution measurements: “packets” of daily observations, that are the number of emergency hospital admissions due
to cardiovascular and respiratory adverse events, are gathered on demand. After anonymization they are securely stored into a Lenvis’ repository and, as for the pollutants measures, made accessible by accounted Lenvis users at any time. In our study we have taken into account the emergency hospital admissions referred to the period from the 1st November 2010 to the 30th April 2011.

2.3 Data cleansing

As a first step of data cleansing we have performed missing values replacement. This issue was relevantly different for the environment and health related data available. In particular, admissions data have been collected on demand and clinical experts have been directly involved in missing values correction. On the contrary, pollutant concentrations are automatically provided by a hardware and software infrastructure which can be affected by failures of devices or in transmission, generating, consequently, missing values. As the sensors network reliability has been significantly improved during latest years, relatively few missing values occurred only for PM$_{2.5}$, that is the most recent measure. In the worst case five consecutive daily measurements of PM$_{2.5}$ are missing, anyway a simple linear interpolation was enough to replace them.

Another crucial data cleansing issue related to the pollutant measurements is the detection of anomalous values (outliers) that can relevantly affect the robustness of analysis. On the contrary, emergency hospital admissions data are not affected by this kind of anomalies, that have been eventually corrected by involving clinical experts as for missing values. On environmental data, we have performed a simple outlier detection-and-removal procedure based on the inter-quartile method: any value $x$ lower than $Q_1-1.5*(Q_3-Q_1)$ or greater than $Q_3+1.5*(Q_3-Q_1)$ is considered as an outlier and removed from the series, where $Q_1$ and $Q_3$ are the lowest and the highest quartile, respectively, of the specific pollutant time series. Since pollutants levels are not distributed normally, but as a Poisson with $\lambda=1$ (i.e. skewness=1.17), a greater number of high values is removed than for the symmetric case. As better specified in the following, the positive asymmetry impacts only on the discretization step by making the forecasting model more “precautionary” with respect to highest concentrations.

2.4 Pollutant-to-pollutant and Pollutant-to-admissions correlations

Preliminary to the definition of the Markov model, we have performed a correlation analysis task (Pearson’s linear correlation) for evaluating the relationship between any pair of pollutants and, then, between each pollutant and emergency hospital admissions in the short term (i.e, at 0 to 5 days), separately for admissions due to respiratory and cardiovascular adverse events.

While it has been pointed out that each pollutant is significantly correlated with each other (p-value < 0.05) - with $O_3$ that showed an inverse correlation - only PM$_{10}$, PM$_{2.5}$ and NO$_x$ provided the highest significant correlations with respiratory emergency hospital admissions (p-value < 0.05), in any time lags, from 0 to 5 days. This result confirms previous findings in literature, such as those reported by Balakrishnan et al. [2011], Barman et al. [2011], Rajarathnam et al. [2011], Sheffield et al. [2011] and Tzivian [2011]. According to the available set of data, weak correlations have been found between pollutant and emergency hospital admissions due to cardiovascular adverse events; respect to this our work focused mainly on the relationship between pollution level and respiratory admissions.

In the following two tables (table 1 and table 2) we report the inter-pollutants and the pollutant versus respiratory emergency hospital admissions correlations.
Table 1. Inter-pollutants correlations

<table>
<thead>
<tr>
<th></th>
<th>NOx</th>
<th>NO2</th>
<th>SO2</th>
<th>O3</th>
<th>CO</th>
<th>Be</th>
<th>PM10</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOx</td>
<td></td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO2</td>
<td>0.35</td>
<td></td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O3</td>
<td>-0.64</td>
<td>-0.61</td>
<td>-0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>0.81</td>
<td>0.66</td>
<td>0.38</td>
<td>-0.56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Be</td>
<td>0.76</td>
<td>0.77</td>
<td>0.40</td>
<td>-0.52</td>
<td>0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM10</td>
<td>0.72</td>
<td>0.69</td>
<td>0.32</td>
<td>-0.46</td>
<td>0.62</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>PM2.5</td>
<td>0.65</td>
<td>0.70</td>
<td>0.36</td>
<td>-0.52</td>
<td>0.64</td>
<td>0.66</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 2. Pollutant to respiratory admissions correlations at different time lags

<table>
<thead>
<tr>
<th></th>
<th>Lag 0</th>
<th>Lag 1</th>
<th>Lag 2</th>
<th>Lag 3</th>
<th>Lag 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOx</td>
<td>0.33</td>
<td>0.21</td>
<td>0.30</td>
<td>0.37</td>
<td>0.36</td>
</tr>
<tr>
<td>NO2</td>
<td>0.30</td>
<td>0.18</td>
<td>0.26</td>
<td>0.36</td>
<td>0.35</td>
</tr>
<tr>
<td>SO2</td>
<td>0.15</td>
<td>0.13</td>
<td>0.11</td>
<td>0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>O3</td>
<td>-0.31</td>
<td>-0.23</td>
<td>-0.27</td>
<td>-0.29</td>
<td>-0.26</td>
</tr>
<tr>
<td>CO</td>
<td>0.32</td>
<td>0.28</td>
<td>0.27</td>
<td>0.28</td>
<td>0.32</td>
</tr>
<tr>
<td>Be</td>
<td>0.23</td>
<td>0.15</td>
<td>0.23</td>
<td>0.30</td>
<td>0.32</td>
</tr>
<tr>
<td>PM10</td>
<td>0.22</td>
<td>0.19</td>
<td>0.28</td>
<td>0.37</td>
<td>0.33</td>
</tr>
<tr>
<td>PM2.5</td>
<td>0.27</td>
<td>0.30</td>
<td>0.32</td>
<td>0.40</td>
<td>0.38</td>
</tr>
</tbody>
</table>

2.5 Data discretization and linear transformation of pollution level

Since Markov models represent processes as states and transitions, a relevant preprocessing has been performed to the discretize pollutants and emergency hospital admissions time series, that are continuous. With respect to the pollutants, we have adopted, as updated guideline, the current European Directive on ambient air quality and cleaner air for Europe that, by law, sets to 50μg/day the limit value for PM$_{10}$ and suggests a lower and an upper threshold, 25μg/day and 35μg/day respectively, to trigger health alert to frail segments of population. Analysis of PM$_{10}$ concentration during winter seasons (November - April) in Milan, for years 2003 to 2011, pointed out that only 43.7% of days, in average, was under the limit threshold mandated by law, while PM$_{10}$ level resulted relevantly lower during summer seasons. In the following Figure 1 PM$_{10}$ daily concentration series for two significant winters are reported: it is easy to note that more days in winter 2008-2009 are under the limit threshold with respect to those in 2002-2003.

![Figure 1. Daily averaged PM10 in two illustrative winter seasons in Milan; limit threshold mandated by law and value used to differ between Severe and Excessive concentrations are the dotted and dashed lines in the figure, respectively.](image_url)
With respect to the discretization we have decided to indicate as “Under Threshold” (U) the daily concentrations of PM\textsubscript{10} lower than the limit threshold and further discretize on two levels the higher ones; we have computed the median value of all the daily concentrations exceeding 50\(\mu\)g/day during all the winters (i.e. 80\(\mu\)g/day) and used it to indicate as “Severe” (S) an “Excessive” (E) the measurements in 50-80\(\mu\)g/day and greater than 80\(\mu\)g/day, respectively. Median value is lower when it is computed after outlier detection-and-removal. This preprocessing is performed only on the historical data used to learn the forecasting model, but not on the streaming data on which forecasts are produced. Respect to this, the learned model will not suppress data as “errors”, instead it will provide more “precautionary” forecasts, by taking into account a lower threshold between Severe and Excessive.

Similarly to the pollutant level, emergency hospital admissions data have been discretized in three different trends: “Stable” (St), “Decrease” (D) and “Increase” (I), according to a variation of 10% with respect to the previous day. This operation produced three different sets having, approximately, the same size.

Plausibly, the influence of pollution on the admissions, at any time lags, depends on the concentration history. We have decided to replace the value of each pollutant in a single day with the weighted average of the values in the last \(n\) days, where the weights are the correlation coefficients of the specific pollutant and the admissions at each time lag. The most suitable number of days \(n\) has been identified for each pollutant that was significantly correlated to respiratory emergency hospital admissions: six for PM\textsubscript{10} and four for PM\textsubscript{2.5} and NO\textsubscript{x}.

3 MARKOV-BASED MODEL

Since pollutants are highly correlated one to each other, a pollutant may be enough to forecast emergency hospital admissions; then, PM\textsubscript{10} level is highly correlated to admissions due to respiratory events and concentration history can be “embedded” by a linear transformation. All these aspects allows us to build the following Markov-based model to forecast hospital admissions due to respiratory events depending on the daily PM\textsubscript{10} concentration during last six days (included today).

The model has two components: (i) a Markov Chain modeling the probable sequences of pollutant concentrations \(C\!=\{C_0, C_1, \ldots, C_t\}\) through a Transition Matrix (T) describing the state evolution process, and (ii) a Conditional Probability Table (CPT) providing the most probable outcome at the time \(t\) according to the status. \(T_{ij}\!=\!P(C_t\!=\!j \mid C_{t-1}\!=\!i)\) is the probability to move from concentration (status) \(i\) to \(j\). Let denote with \(O\) the CPT: this matrix has one column for each state of the Markov Chain and the entry \(O_{ij}\!=\!P(O_t\!=\!i \mid C_t\!=\!j)\) is the probability to observe trend \(i\) (St, D or I) when model is in the state \(j\), as described in Russel et al. [2003]. Although this definition is similar to that of a Hidden Markov model, in our approach the state is known and not hidden. Parameters to be learned are the entries of \(T\) and \(O\): according to the normalization condition, six parameters for \(T\) and six for \(O\) have been learned, by using Maximum Likelihood.
RESULTS

Initially, only the Markov Chain has been trained on the data related to PM$_{10}$ during the winter seasons 2002-2003 to 2010-2011. By using the Markov Chain we have computed the stationary distributions for each winter, that are the probabilities to persist in each one of the possible state; results are shown in Figure 3.

Probability to observe a long series of days having daily PM$_{10}$ concentration Under Threshold (i.e., under limit threshold mandated by law) increased relevantly from 2007: in that year the “Ecopass” pollution control policy was brought into force, aimed at restricting the vehicles access to the center of Milan. A slow decrease in probability can be noted after two years, when Ecopass was relaxed.

Another relevant analysis that we performed through the only Markov Chain is the Maximum Run Length. The aim was to provide the environment authority with an objective evaluation of the air quality. Result of this analysis is the probability to persist $r$ consecutive days, for each pollution level. In Figure 4 the Maximum Run Length for winter season 2010 is depicted. Probability to persist in a Severe status decreases, with $r$, faster than for Excessive and Under Threshold; therefore, probability to have more than three days with an Under Threshold PM$_{10}$ level was relevant higher than for Severe and Excessive during the analyzed period.
Finally, we have computed the Transition Matrix and the Conditional Probability Table by using pollutants and admissions data both related to winter 2010 and build a complete model similar to that in Figure 2, with the aim to forecast respiratory admissions trend at time lag $k$ depending on PM$_{10}$ concentration. To estimate the reliability of the model we have split the dataset in two: a portion to train the model (November 2010 to March 2011, 151 days) and one to test it (April 2011, 27 days). We have used time lag $k=4$ (for which correlation is the highest) and daily PM$_{10}$: recall for “Increase” was really poor (27%) but model proved to be effectively useful to predict “Decrease” trends (recall=80%) supporting efficient resources planning in the emergency hospital settings (costs reduction).

5 CONCLUSIONS

The Markov-based analysis proposed in this paper proved to be useful to develop decision support services for environment and healthcare stakeholders. In detail, we have shown how to adopt a Markov-based model to provide objective evaluation of both global quality of air during a given period of time and pollution control policies effectiveness, in particular the Ecopass policy brought into force in 2007 in Milan. Respect to other studies focused on specific zones Milan and for a really limited (3 days) time window, such as Invernizzi et al [2011a][2011b], we have used a large dataset of daily pollutants time series (2002 to 2010) that essentially provide robustness to the results about the effectiveness of Ecopass in reducing PM$_{10}$.

With respect to the admissions forecasting service, it can be used by hospital managers in order to plan cost-rational planning of resources (e.g., by reducing or increasing them according to the predicted trend). Also General Practitioners may use it with the aim to assess/predict the health risk of a patient and prevent hospitalization by prompt and preventive actions, reducing costs while improving health/living quality, as well as plan outdoor activities limiting risks due to amplified respiration, induce for example by exercising, as reported by Sharman et al. [2004], with consequent lung deposition and pollutants inhalation.

REFERENCES


Zeger, S.L., Dominici, F., Samet, J., Harvesting-resistant estimates of air pollution effects on mortality, Epidemiology, 10(2), 171-175, 1999.