CHEMICAL ENGINEERING TRANSACTIONS

VOL. 74, 2019

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Sauro Pierucci, Jiří Jaromír Klemeš, Laura Piazza
Copyright © 2019, AIDIC Servizi S.r.l.
ISBN 978-88-95608-71-6; **ISSN** 2283-9216

DOI: 10.3303/CET1974046

A Bayesian Belief Network for Local Air Quality Forecasting

Tomaso Vairo[ac]*, Mario Lecca[a], Elisabetta Trovatore[a], Andrea P. Reverberi[b], Bruno Fabiano[c]

[a] ARPAL, via Bombrini 8 - 16149 Genoa Italy
[b] DCCI - Chemistry and Industrial Chemistry Dept., Genoa University, via Dodecaneso 31 - 16145 Genoa, Italy
[c] DICCA - Civil, Chemical and Environmental Engineering Dept. – Genoa University, via Opera Pia 15 - 16145 Genoa, Italy
tomaso.vairo@arpal.gov.it

This study is focused on the development of a Bayesian network for air quality assessment and aims at offering a pragmatic and scientifically credible approach to modelling complex systems where substantial uncertainties exist. In particular, the main object is the prediction of the occurrence of suitable conditions for the stagnation of pollutants in a given area. The analytical modeling of the network provides a set of independent nodes, represented by the outputs of a forecasting meteorological Limited Area Model, from which descend the conditions for the stagnation of pollutants in different areas of the city (through measurements of the heuristic pollutant from monitoring stations) and finally the global conditions. The urban area of Genoa (Italy) was selected in order to test the actual capability of the model prototype. Network training was performed by means of historical data resulting from significant statistical series of the past years by the air quality-monitoring network. The system used for data assimilation, construction and network learning is completely based on an open source statistical processing software.

1. Introduction

Atmospheric pollution in urban areas has become one of the key environmental issues to be faced, given its serious adverse effects on human health. Two main modelling approaches can be outlined: the former is based on numerical modelling by simulating atmospheric dispersion and transport starting from emission source characterization. It is widely applied also when dealing with accidental hazardous releases, risk areas evaluation (Palazzi et al., 2004) and possible delayed ignition in case of flammable materials (Pesce et al., 2012). The latter is based on advanced statistical models including artificial intelligence, machine learning technologies and soft computing techniques (e.g. Alimissis et al., 2018). As widely acknowledged, the quantification of uncertainty of modelling results and the optimal selection of the atmospheric dispersion model relies on an accurate evaluation on how model uncertainty in model inputs affects the outputs (Vairo et al. 2014). In case of complex geometries, such as urban areas additional limitations arise, so that, for example, the numerical or analytical resolution of the relevant inverse problem starting from field experimental data back to the pollution source requires a proper regularization technique (e.g. Reverberi et al., 2013). The choice of appropriate level of detail and complexity is based on the purpose and application area the model is to be used (Baklanov et al., 2008). The focus of this work is to evaluate the results that the Bayesian networks (BN) could achieve to perform a reliable forecasting of critical air quality conditions, connected to the stagnation of pollutants. We have tested several different configurations in order to investigate the relevance of different factors, such as the net topology, availability and consistency of historical data, evidence provided and so on. Bayesian networks automatically capture probabilistic information from data using directed acyclic graphs and factorized probability functions. The graph intuitively represents the relevant spatial dependencies and the resulting factorized probability function leads to efficient inference algorithms for updating probabilities when new evidence is available. The remainder of this paper firstly outlines the development of Bayesian networks using two approaches, i.e. self-construction from experimental air pollution database and critical BN development by application of expert knowledge, using both discrete and continuous (Gaussian) nodes. Subsequently, the network is tested with different learning algorithms (hill climbing, Grow-Shrink, Incremental

Paper Received: 28 July 2018; Revised: 28 January 2019; Accepted: 15 May 2019

Please cite this article as: Vairo T., Lecca M., Trovatore E., Reverberi A., Fabiano B., 2019, A Bayesian Belief Network for Local Air Quality Forecasting, Chemical Engineering Transactions, 74, 271-276 DOI:10.3303/CET1974046

Association, Tabu search), Predictions and forecasting methods, including their estimation in an iterative Markov Chain Monte Carlo (MCMC) computation setup. At last, the results of the network accuracy based on the scores are represented and critically discussed in the perspective of providing an efficient and robust option for air quality forecasting at the local scale.

2. Analytical modelling and Bayesian Reasoning

Analytical models are mathematical models that have a closed-form solution, i.e. the solution to the equations used to describe changes in a system can be expressed by a mathematical analytical function (e.g. Aguilera et al., 2011). The purpose of the analysis is to understand something about the nature of the real world by studying the observed data in relation to their context. Figure 1 depicts a schematic view of analytical modelling: X-axis reflects the purpose of modelling, ranging from association/correlation to causality. Y- axis represents the source of the model conceptually ranging from theory to data. The theory (parametric) is labelled as Human Intelligence, suggesting the origin. At the opposite end of the Y-axis, data are commonly associated with Machine Learning and Artificial Intelligence.

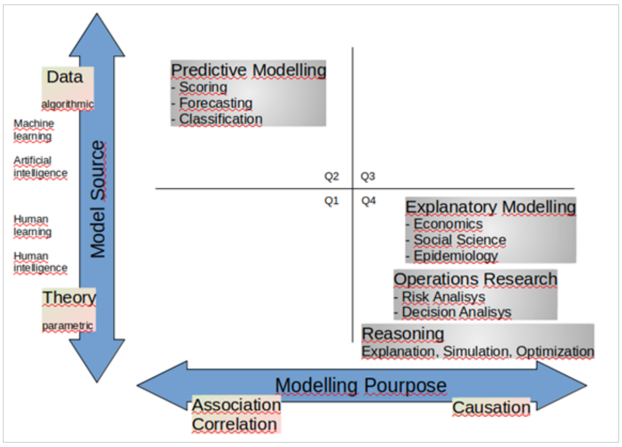


Figure 1: Map of analytical modelling.

The Bayesian network, also known as the reliability network, is an extension of the Bayesian method and is one of the most effective theoretical approaches for uncertain knowledge representation and reasoning, which allows analysing the phenomena accepting evidence for any node and therefore using the Bayes theorem to calculate the a posteriori probabilities of the events. As for the conceptual division between theory and data in Fig. 1, a dynamic network can be based on expert knowledge, i.e., from the theory (Human Learning), or it can be predictive, i.e. the machine learns from the data (Machine Learning). In this respect, dynamic networks can use the entire spectrum as the source of the model. As amply discussed in different contexts (Pasman and Rogers, 2013), the basic concept of the bayesian approach regards the conditional probability that indicates the probability of the event *A* given the event *B*. The Bayesian networks allow us to represent the joint probability distribution for the environment of interest and their name descends from the Bayes theorem that is the fundamental mathematical tool in order to update the values of the networks when they are subjected to some new evidences according to Eq.(1):

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{1}$$

where:

P(A) is the prior probability (or “unconditional” or “marginal” probability) of *A*.

P(A|B) is the conditional (posterior) probability of *A*, given *B*. *P(B|A)* is the conditional probability (likelihood) of *B* given *A*. *P(B*) is the prior or marginal probability of B, and acts as a normalizing constant.

$\frac{P(B|A)}{P(B)}$ is the Bayes factor or likelihood ratio.

The Bayesian approach is based upon the probability assigned to an event as a consequence of the current knowledge (inference process). A Bayesian network provides a complete description of the application domain in the form of a conditional probability distribution and is a directed acyclic graph in which each link is directed from a parent node to its child. Each node represents a variable of the domain and the links represent causal dependencies among the variables. The possible values of each variable are divided into some intervals (of a given width), or are considered in a Gaussian distribution. Evidence is called the a-priori information about the degree of certainty assigned to the possible states of a variable. Each variable without parents is characterized by an a-priori probability table, while each variable A with some parents $B_1,...,B_n$ has a conditional probability table that expresses the joint probability. A generic element of the conditional probability table is the probability that occurs as a combination of two specific values of the variables. More in general, we can express it in the form: $P(X_1 = x_1 / ... / X_n = x_n)$ and consequently obtain Eq. (2).

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P\left(x_i|Parents(X_i)\right) \tag{2}$$

The construction of a Bayesian network can be performed according to the following phases:

1. A set of variables *Ai (i=1, ..., n)* is chosen to describe the system;
2. To each variable *Ai* is associated a node of the network;
3. The set *Parents (Ai)* of the nodes parents of *Ai* is fixed;
4. The table of the conditioned probabilities for *Ai* is defined by the learning of a set of cases.

3. The Bayesian model

The here defined Bayesian Beliefs Network for Air Quality (BBNAQ) system relies on following data acquisition and elaboration:

- Forecasting data of temperature *T*, wind *W*, humidity *H* and rain *R*, from MOLOCH meteorological model (2013-2016) as described below.
- PM10 concentration data (2013-2016) from three stations in the urban monitoring network of Genova (*S1, S2, S3*), as daily average concentration.
- The data are inserted in an analytical model which is developed ad-hoc starting from the the R statistical processing approach (Scutari, 2010).

The meteorological data were extracted from the archive of the operational meteorological model MOLOCH, which provides two-days high resolution forecast. It is a limited area model, by ISAC-CNR (Italy), which integrates the non-hydrostatic, fully compressible equations for the atmosphere, with 50 atmospheric levels and 7 soil levels, with a resolution of 0.02°, corresponding roughly to 2 km. The forecasting data were extracted from the previous day run and consist on the following parameters: wind intensity at 10 meters, temperature at 2 meters, relative humidity at 2 meters and total precipitation cumulated in the previous 3 hours. The logic steps of the analytical modeling can be summarized as follows:

- data extraction from MOLOCH, processing and labeling suitable for the construction of the network;
- collection of concentration data from the control units, processing and labeling suitable for the construction of the network;
- preliminary statistical analysis;
- development of the network with different learning algorithms (hill climbing, Grow-Shrink, Incremental Association, Taboo search);
- validation of the consistency of the network.

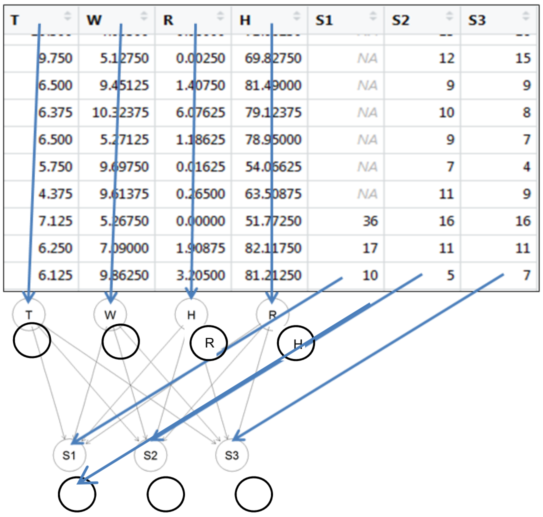
Learning is based on *T, W, H, R* and *S1, S2, S3* historical data recorded starting from the data 2013-2016. It is noteworthy noting that all the sampling stations (*S1*-*S3*) are located within the urban center of Genoa. *S3* sampling station is positioned between some high bulkheads, so that it is expected no correlation with the intensity of the wind. A two-stage modelling approach was adopted: the Bayesian network was manually established by expert knowledge, and then the previously obtained Bayesian network model is corrected by learning the data sets suitable to perform an automatic construction of BBN.

3.1 BBN from expert knowledge

In this case, an empty DAG (Direct Acyclic Graph), i.e. an empty network with dependencies decided by theory and knowledge was created. Subsequently, the fitting between the dataset and the DAG was performed (i.e. in the "*T*" node, all the Temperature data are fitted) and the same for other nodes, as shown in Figure 2). As discussed in the following, a subset of available data was subsequently utilized for validation of the predictive ability of the scheme. Two methods were then tested: the first, dividing the data, in each node, in 5 discrete intervals (discrete value network depicted in Figure 3a), the second, leaving the data according to a continuous distribution (Gaussian network – Figure 3b).

3.2 Automatically built BBN

In this case, the building of the network with continuous values (Gaussian Bayesian network) is directly derived from the whole data set. The structure of the network derives from the data analytics (clusters analysis and correlations). As depicted in Figure 4, the automatically built structure evidences analytical dependencies between the nodes that are often not intuitive, but that, with the use and validation of the model, are proved statistically significant. A noteworthy example in the case study is the attainment of the independence of station *S3* from the parameter "wind", connected to the actual localization of this monitoring station in a sheltered position, preventing wind action. This evidence was correctly captured by the automatic construction of the network but from the “expert knowledge” analysis, based only on raw data acquisition. In the next section, we discuss the results, critically comparing the different approaches previously outlined. The response is obtained using the MCMC (Markov Chain Monte-Carlo) sampling method.



T	W	R	H	S1	S2	S3
9.750	5.12750	0.00250	69.82750	NA	12	15
6.500	9.45125	1.40750	81.49000	NA	9	9
6.375	10.32375	6.07625	79.12375	NA	10	8
6.500	5.27125	1.18625	78.95000	NA	9	7
5.750	9.69750	0.01625	54.06625	NA	7	4
4.375	9.61375	0.26500	63.50875	NA	11	9
7.125	5.26750	0.00000	51.77250	36	16	16
6.250	7.09000	1.90875	82.11750	17	11	11
6.125	9.86250	3.20500	81.21250	10	5	7

Figure 2: BBN from expert knowledge. Each node contains the proper value derived from the whole dataset.

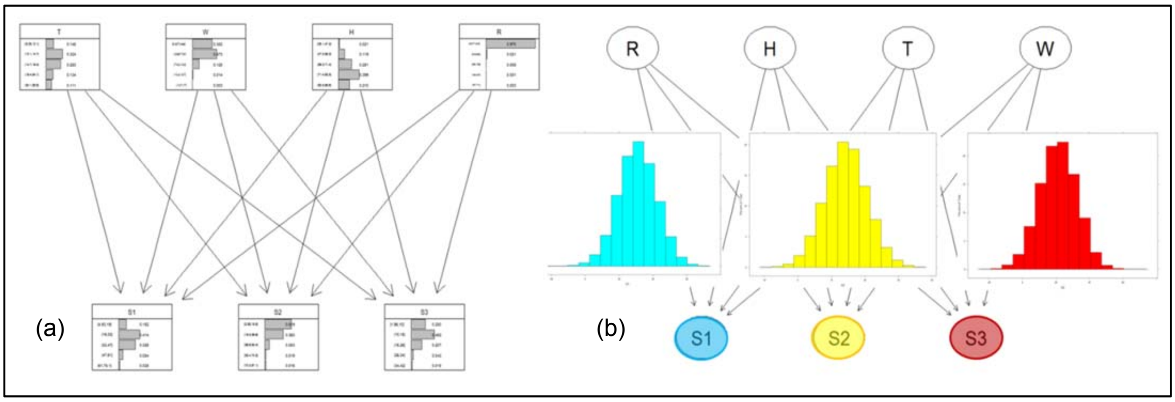


Figure 3: (a) Discrete network from expert knowledge: in each node, the prior probabilities are reported, where the data in the dataset are divided into 5 discrete intervals. (b) Gaussian Network from expert knowledge: in each node, the frequency distribution is reported.

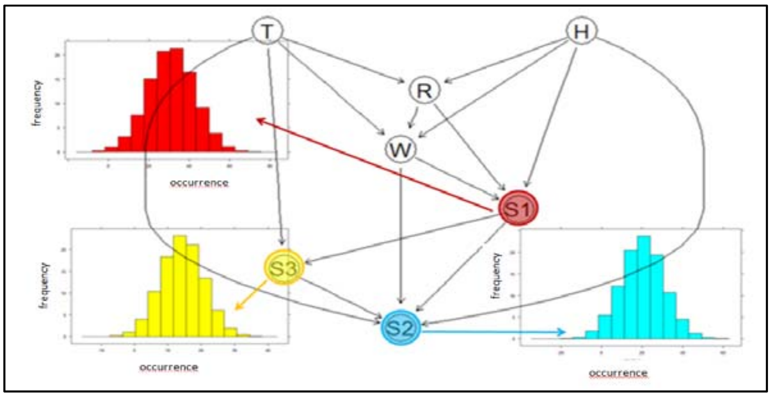


Figure 4: Gaussian network obtained from data.

4. Results

Figure 5 depicts the results obtained by the network with continuous values (Gaussian Bayesian network) – from expert knowledge, whose distribution parameters are summarized in Table 1.

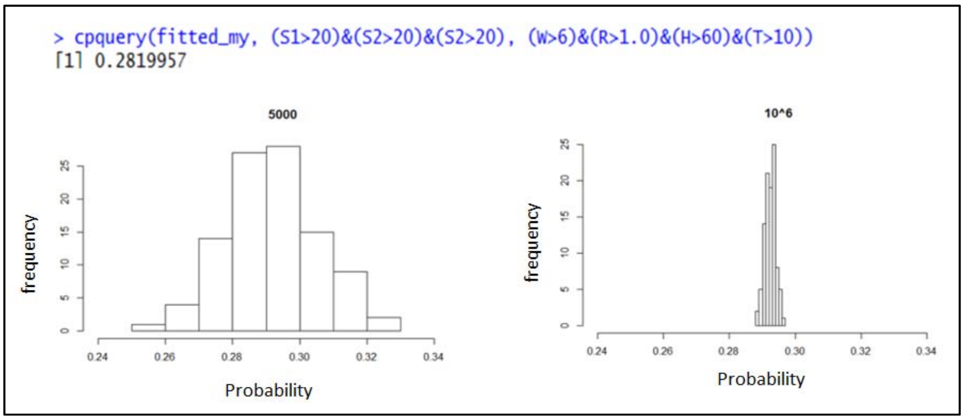
Table 1: Gaussian BBN from expert knowledge - distribution parameters.

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
0.2792	0.2890	0.2926	0.2927	0.2955	0.3125

Table 2 and Figsures 6 a-b provide comparison between the prediction ability of BBN from expert knowledge or from data analytics. The last approach allows a free implementation of structure learning algorithms along with the conditional independence tests and network scores (Scutari, 2010). Statistical performance of each scheme are evaluated by significant correlation measures (i.e. Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE)). As shown in Figs. 6(a) and 6(b) comparing model predictions and observed values, the best configuration among the tested modelling approaches is given by the automatically built from data analytics Gaussian network.

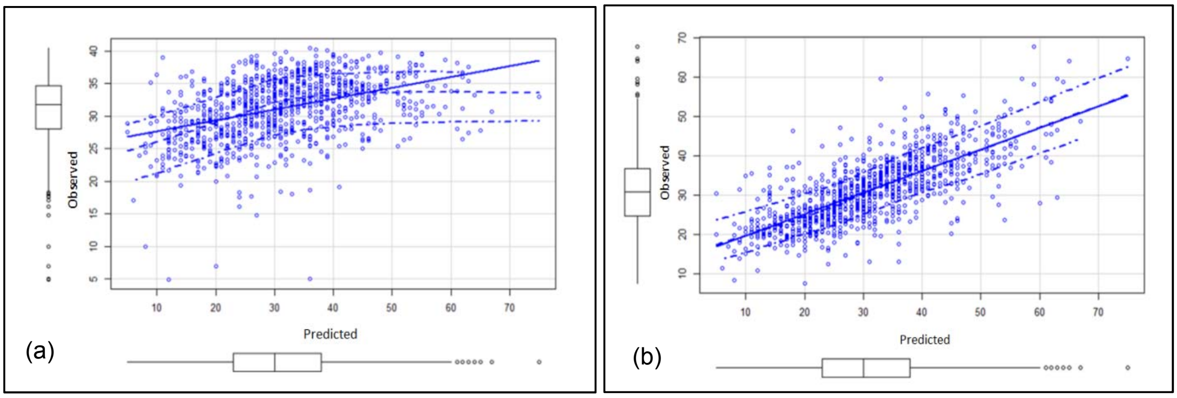
Table 2: Gaussian BBN observed vs. predicted - distribution parameters: expert knowledge vs. data analytics.

	ME	RMSE	MAE	MPE	MAPE
Expert knowledge	2.24514e-14	10.45438	8.219454	-14.71306	33.29295
Data analytics	0.0214116	7.679559	5.761562	- 8.654668	23.18986



```
> cpquery(fitted_my, (S1>20)&(S2>20)&(S2>20), (W>6)&(R>1.0)&(H>60)&(T>10))
[1] 0.2819957
```

Figure 5: Gaussian BBN from expert knowledge: the accuracy of the result depends upon the actual number of iterations in Markov Chain Monte-Carlo*.*



Figures 6 (a) and (b): (a) Gaussian BBN from expert knowledge. (b) Gaussian BBN from data analytics. Observed vs. predicted.

As shown in Figures 6(a) and 6(b) comparing model predictions and observed values, the best configuration among the tested modelling approaches is given by the automatically built from data analytics Gaussian network. As clearly indicated also by the statistical parameters summarized in Table 2, the BBN from data analytics shows a better prediction ability. MAE represents the sum of two components: Quantity Disagreement and Allocation Disagreement. Even if the BBN from expert knowledge evidences a better Quantity disagreement (ME), the BBN from data analytics shows a far better allocation disagreement, globally allowing an improvement of all the distribution parameters utilized, which from an applicative viewpoint translates into a better predictive ability.

5. Conclusions

Currently, the proposed framework is under experimental testing at the meteo-hydrological centre of the Liguria region, in order to daily issue an air quality bulletin. From a methodological point of view, the most important results obtained from the tested network models can be summarized in the following points:

- the Gaussian network provides better and more reliable forecasts than the discrete network;
- the network built automatically by the data provides better results of the network built by expert knowledge;
- the optimal configuration in the given urban context corresponds to the self-built Gaussian network.

Further developments of the system currently under investigation include scare validation of the developed network, a new network implementation with incoming PM_{10} concentration data from the previous day as well as a cross check comparison obtained by a non-Bayesian neural network (i.e., by Error back-propagation).

Acknowledgments

The authors would like to acknowledge all the staff of the meteo-hydrological centre of environmental protection agency of Liguria (ARPAL – CFMI), as well as prof. Marco Scutari, of Oxford University, UK.

References

Alimissis A., Philippopoulos K., Tzanis C.G., Deligiorgi D., 2018, Spatial estimation of urban air pollution with the use of artificial neural network models, Atmospheric Environment 191, 205-213.

Aguilera P.A., Fernández A., Fernández R., Rumí R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environmental Modelling & Software 26, 1375-1774.

Baklanov, A., Korsholm, U., Mahura, A., Petersen, C., and Gross, A., 2008. ENVIRO-HIRLAM: on-line coupled modelling of urban meteorology and air pollution, Adv. Sci. Res., 2, 41-46

Pasman H.J, Rogers W.J., 2013, Bayesian networks make LOPA more effective, QRA more transparent and flexible, and thus safety more definable! Journal of Loss Prevention in the Process Industries, 26, 434-442.

Palazzi E., Currò F., Fabiano B., 2004, Simplified modelling for risk assessment of hydrocarbon spills in port area, Process Safety and Environmental Protection 82, 412-420.

Pesce, M., Paci, P., Garrone, S., Pastorino R., Fabiano, B., 2012, Modelling ignition probabilities within the framework of quantitative risk assessments. Chemical Engineering Transactions 26, 141-146.

Reverberi A.P., Fabiano B., Dovì V.G., 2013, Use of inverse modelling techniques for the estimation of heat transfer coefficients to fluids in cylindrical conduits. Int. Commun. in Heat and Mass Transfer 42, 25-31.

Scutari M., 2010. Learning Bayesian Networks with the bnlearn R Package. J. of Statistical Software 35, 1-22.

Vairo T., Currò F., Scarselli, S., Fabiano B., 2014, Atmospheric emissions from a fossil fuel power station: dispersion models comparison. Chemical Engineering Transactions 36, 295-300.