Информационные процессы, Том 9, № 4, 2009, стр. 280–286.

INFORMATION PROCESSES

Ordinal Data Mining for Fine Particles with Non Parametric Continuous Bayesian Belief Nets

A. Hanea* **and W. Harrington****

**Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands*
a.hanea@ewi.tudelft.nl
***Resources for the Future, Washington DC, United States*
harrington@rff.org
Received 30.10.2009

Аннотация—We introduce a Bayesian Belief Net (BBN) based approach for analysing the relationship between SO_2 emissions, and concentrations of fine particulate matter, $PM_{2.5}$. $PM_{2.5}$ exposure has adverse health effects, hence we study this relationship with the goal of quantifying the health benefits of emission reductions. The main advantage of our approach is that it can handle a large number of continuous variables, without making any assumptions about their marginal distributions, in a fast manner. Rapid computations are of little value if the model itself cannot be validated. We discuss the issue of validation and additionally perform backward and forward inference.

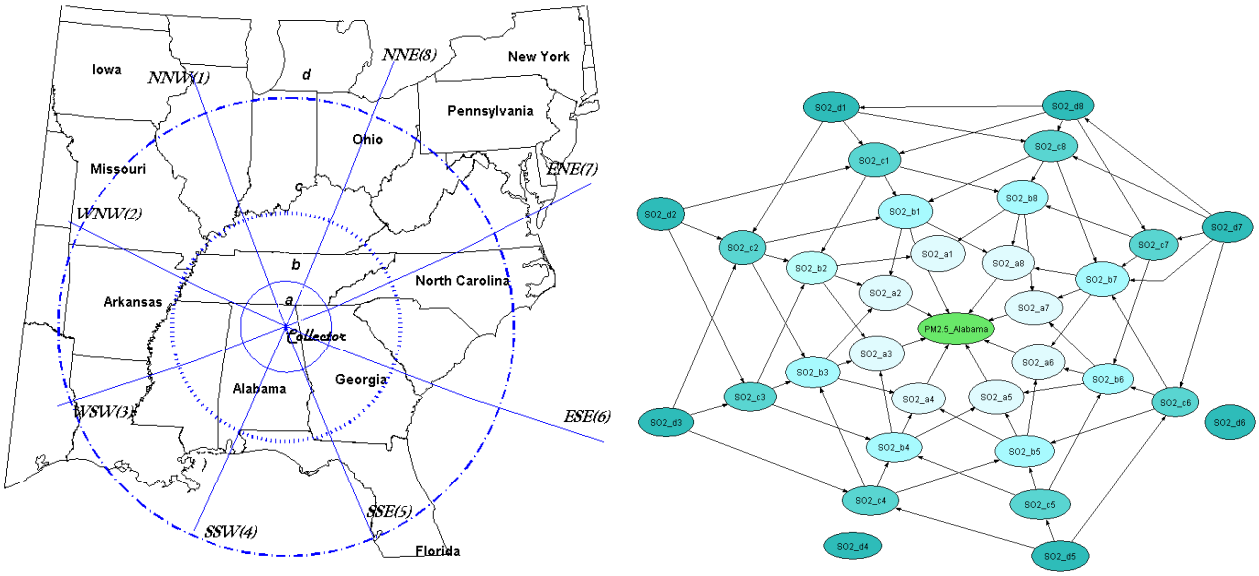
1. INTRODUCTION

This paper introduces a non parametric Bayesian belief net (NPBBN) based approach for analysing the relationship between SO_2 emissions from power plants, and ambient concentrations of fine particulate matter $PM_{2.5}$. $PM_{2.5}$ exposure has adverse health effects for humans, hence we study this relationship with the final goal of quantifying the health benefits/risks of emission reductions/increase.

The full data set comprises monthly emissions of SO_2 gathered from electricity generating stations, and monthly mean concentrations of $PM_{2.5}$ gathered from collection sites in the United States over the course of seven years (1999 – 2005)[1]. We have 84 monthly averages of emitters and collectors. For each monitoring site we consider 32 variables corresponding to SO_2 emission data from its vicinity. These 32 variables are denoted SO2_a1, SO2_a2, ..., SO2_d8, depending on the direction and distance from the monitoring site. Ring "a" consists of power plants within 100 miles of the monitor site, ring "b" to 100 – 250 miles, ring "c" corresponds to 250 – 500 miles, and ring "d" contains all plants further than 500 miles from the monitor. The numbers correspond to 8 zones whose bisectors are the compass directions NNW, proceeding counterclockwise to NNE. We focus on relative spatial relationships of sources and receptors, not on their absolute locations, hence for building a NPBBN and exploring its properties the location of the monitor was arbitrarily chosen to be a rural site located about 30 km north of Gadsden, in the state of Alabama. Figure 1a shows a map of the area in focus.

Definitions and concepts are introduced in Section 2, but suffice to say now that NPBBNs are directed acyclic graphs (DAGs) where an arrow connecting a parent node to a child node indicates that influence flows from parent to child. A NPBBN for Alabama ambient $PM_{2.5}$ is shown in Figure 1b. We are interested in the dependence between the $PM_{2.5}$ concentration PM2.5_Alabama and the emissions, and the interdependence among emissions. Some of these relationships are represented

[1] Assembly of the data set was supported by a grant from the Health Effects Institute.



(a) Selected collector and area around it. (b) Variables selected for the NPBBN model.

Fig. 1. Selected area, collector site and emission stations.

by arcs in the DAG from Figure 1b.

The dependence (measured as correlation) between an emitter and the collector might be modest, owing to other factors influencing $PM_{2.5}$ concentrations. At the same time the correlations between the emitters themselves are rather strong, as they are associated strongly with a common driver, the weather. In addition, the number of data points is small. These factors lead to the conclusion that a standard linear regression model might encounter difficulties and result in biased estimates of regression coefficients. We propose an alternative method for analysing emitter-collector data based on NPBBNs.

2. NPBBNS: LEARNING AND VALIDATION

Bayesian belief nets (BBNs) are DAGs, whose nodes represent univariate random variables, which can be discrete or continuous. The arcs represent direct influences. BBNs provide a compact representation of high dimensional distributions of a set of variables and encode their joint density/mass function by specifying a set of conditional independence statements (in a DAG form) and a set of probability functions.

Until recently, BBNs were discrete, normal or discrete-normal (e.g. [1]). Despite their popularity, they suffer from severe limitations. Discrete BBNs are limited by size and complexity, normal and discrete-normal BBNs are limited by the assumption of joint normality. NPBBNs were introduced to overcome the above mentioned limitations (e.g. [2]).

A *NPBBN* is a DAG, together with a set of (conditional) rank correlations, a copula that represents independence as zero correlation, and a set of marginal distributions.

In NPBBNs nodes are associated with arbitrary distributions and arcs with (conditional) rank correlations that are realised by a chosen copula. Therefore, every arc in the DAG is assigned a (conditional) rank correlation between parent and child. These assignments together with the DAG, the choice of the copula, and the marginals uniquely determine the joint distribution [3]. The

(conditional) rank correlations assigned to the edges of a NPBBN are algebraically independent. The dependence structure is meaningful for any such quantification, and need not be revised if the univariate distributions are changed.

In continuous NPBBNs, nodes are associated with continuous invertible distribution functions. In this paper the nodes of a NPBBN will be assumed continuous. In principle any copula can be used in building a NPBBN, but in practice, only the normal copula enables rapid computations required for large problems. Hence our procedure uses the normal copula.

The distinctive feature of learning a NPBBN from a data set is that the one-dimensional marginal distributions are taken directly from data, and the model assumes only that the joint distribution has a normal copula. That is to say that the variables' rank dependence structure is that of a joint normal distribution.

The concepts of learning and validation are closely connected, as indeed the goal is to learn a NPBBN that is valid. Validation involves two steps: validating that the joint normal copula adequately represents the multivariate data, and validating that the NPBBN is an adequate model of the saturated graph.

Testing the normal copula hypothesis depends on the fact that the rank correlation of two standard normal variables Z_i and Z_j is given by the Pearson transformation of the product moment correlation [5]:

$$\frac{6}{\pi}\arcsin(\frac{\rho(Z_i,Z_j)}{2}) = r(Z_i,Z_j) = \rho(\Phi(Z_i),\Phi(Z_j))$$

where r denotes the rank correlation, and ρ denotes the product moment correlation. The cumulative distribution function of Z_i, applied to Z_i, $\Phi(Z_i)$, returns the rank of Z_i and is uniformly distributed. To determine the empirical rank correlation of a set of bivariate observations of X_i and X_j, we first construct the empirical distribution functions $F_{X_i}(q)$ and $F_{X_j}(q)$. The empirical rank correlation of two variables X_i and X_j is given by:

$$r(X_i,X_j) = \rho(F_{X_i}(X_i),F_{X_j}(X_j)) \tag{1}$$

This involves no modelling hypotheses, and is simply computed from data.

The empirical normal version of X_i is defined as:

$$Z_i = \Phi^{-1}(F_{X_i}(X_i)) \tag{2}$$

Applying Pearson's transformation, we find the empirical normal rank correlation of X_i and X_j to be:

$$\frac{6}{\pi}\arcsin(\frac{\rho(Z_i,Z_j)}{2}). \tag{3}$$

This relation is characteristic to the normal distribution. If X_i and X_j are joined by the normal copula, the empirical rank correlation will approach the empirical normal rank correlation as the number of observations becomes large. Otherwise it will not.

Validation requires an overall measure of multivariate dependence on which statistical tests can be based. A suitable measure in this case is the determinant of the rank correlation matrix [4]. The determinant is 1 if all variables are independent, and 0 if there is linear dependence between the

normal versions of the variables.

For the first validation step we distinguish 2 determinants. DER is the determinant of the empirical rank correlation matrix. DNR is the determinant of the empirical normal rank correlation matrix[1]. DNR will generally differ from DER because DNR assumes the normal copula, which may differ from the empirical copula. A statistical test for the suitability of DNR for representing DER is to obtain the sampling distribution of DNR and check whether DER is within the 90% central confidence band of DNR. One can do that as follows:

1. Compute DER using Equation (1)
2. Construct the normal version of each variable using Equation (2) and calculate the product moment correlation matrix
3. Compute DNR using Equation (3)
4. Re-sample the "normal" data to obtain the distribution of DNR as follows:
   (a) If there are n variables in the original data file, sample from an n-dimensional normal distribution with standard normal margins and the product moment correlation matrix calculated in step 2)
   (b) Compute the determinant of the rank correlation matrix using Equation (1)
   (c) Repeat steps a) and b) 1000 times
   (d) Extract the 5th and 95th quantiles of the distribution of the determinant obtained in step c)
5. Compare DER with the bounds in step 4d). If DER is within these bounds, do not reject the joint normal copula, otherwise reject.

If the normal copula assumption is not rejected on the basis of this test, we shall attempt to build a NPBBN which represents the DNR parsimoniously. Note that the saturated DAG will induce a joint distribution whose rank determinant is equal to DNR, since the NPBBN uses the normal copula. However, many of the influences only reflect sample jitter and we will eliminate them from the model.

One can calculate from data a conditional rank correlation specification for the arcs of a NPBBN. Using the normal copula, the conditional rank correlations can be transformed to conditional product moment correlations. For normal variables, conditional product moment correlations are equal to partial correlations. Hence we can translate a conditional rank correlation specification into a partial correlation specification.

The following property of NPBBNs gives the main reason for using the determinant of the correlation matrix as a measure of multivariate dependence [4].

Theorem 1. *Let D be the determinant of an n-dimensional correlation matrix ($D > 0$). For any partial correlation NPBBN specification*

$$D = \prod \left(1 - \rho^2_{ij;D_{ij}}\right), \tag{4}$$

where $\rho_{ij;D_{ij}}$ is the partial correlation associated with the arc between node i and node j, D_{ij} is the conditioning set for the arc between node i and node j, and the product is taken over all arcs in the NPBBN.

In other words, a NPBBN is a way of factorising the determinant of the correlation matrix. Small partial correlations do not significantly contribute to the product in Equation (4). The value of the

[1] That is the rank correlation matrix obtained by transforming the marginals to standard normals, and then transforming the product moment correlations to rank correlations using Pearson's transformation.

determinant of the correlation matrix is driven by the largest partial correlations. Moreover, the zero partial correlations will correspond to the conditional independence statements encoded in the DAG structure.

Let us return to our learning procedure. We will build the NPBBN by adding arcs between variables only if their rank correlation is among the largest. The heuristic we are using is that partial correlations are approximately equal to conditional rank correlations. This is a reasonable approximation for the normal copula. The procedure for building a NPBBN to represent a given data set is not fully automated, as the directionality of (some of) the arcs will reflect causal or temporal relations which can never be extracted from data. The result of introducing arcs to capture causal or temporal relations is called a skeletal BBN. Let DBBN denote the determinant of the rank correlation matrix of a NPBBN using the normal copula. The second validation step is similar to the first.

1. Construct a skeletal NPBBN
2. Re-sample to obtain the distribution of DBBN (use a procedure similar to the one in the first validation step)
3. If DNR is within the 90% central confidence band of DBBN, then stop, else continue with the following steps;
4. Find the pair of variables such that the arc between them is not in the DAG and their rank correlation is greater than the rank correlation of any other pair not in the DAG. Add an arc between them and recompute DBBN together with its 90% central confidence band;
5. If DNR is within the 90% central confidence band of DBBN, then stop, else repeat step 4).

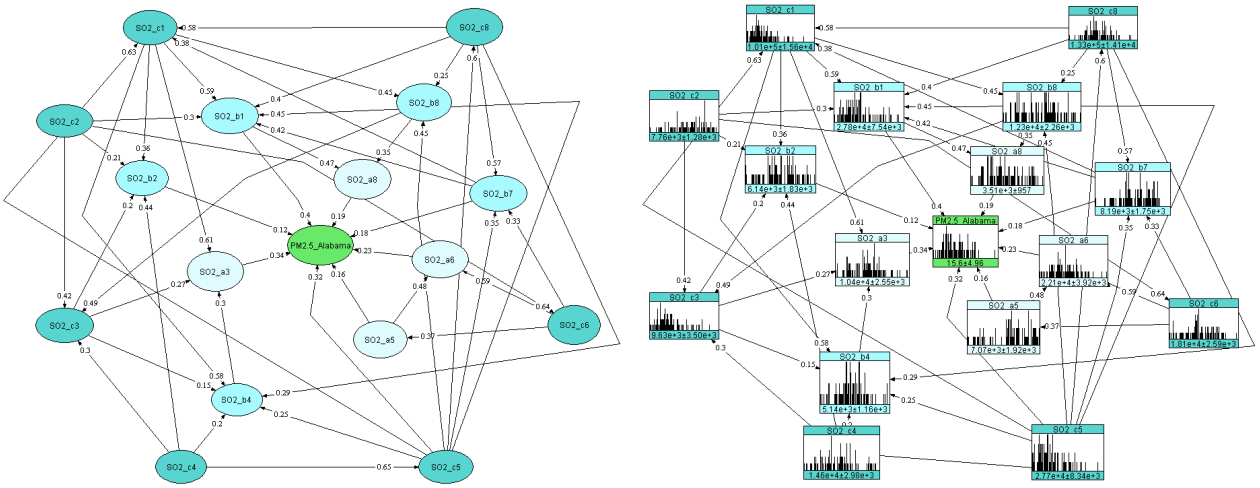
The resultant NPBBN may contain nodes that have more than one parent. If the correlations between the parents of a node are neglected in the NPBBN (i.e. if the parents are considered independent), then DBBN will be different for different orderings of the parents. These differences will be small if the neglected correlations are also small. In general, there is no "best" model; the choice of directionality may be made on the basis of non-statistical reasoning. Some influences may be included because the user wants to see these influences, even though they are small. There may be several distinct NPBBNs which approximate the saturated NPBBN equally well.

3. LINKING $PM_{2.5}$ CONCENTRATIONS TO SO_2 EMISSIONS

We illustrate the method described in Section 2 using only a subset of the emitter-collector data presented in Section 1. The approach is designed to handle a very large number of variables, but a smaller number is more suitable for explaining the method. We choose 17 from the 33 variables showed in Figure 1b. We exclude the variables that correspond to areas where no power plants exist and those corresponding to ring "d", since emissions close to receptors tend to have stronger effects than emissions further away. Finally another small number of variables is excluded such that the distribution of the locations of remaining power plants is "uniform" in all directions. The 17 variables left in the model are shown in Figure 2. So are their marginal distributions. Means and standard deviations are displayed under the histograms[3].

Before starting to build a model we test the hypothesis that the data were generated from the joint normal copula. Following the procedure described in Section 2 we calculate DER = 0.495E-05 and DNR = 1.326E-05. Based on 1000 simulations, we determine that the 90% central confidence band for DNR is [0.037E-05, 1.695E-05]. The hypothesis that the data were generated from the joint normal copula would not be rejected at the 5% level. We start building a model that will reconstruct

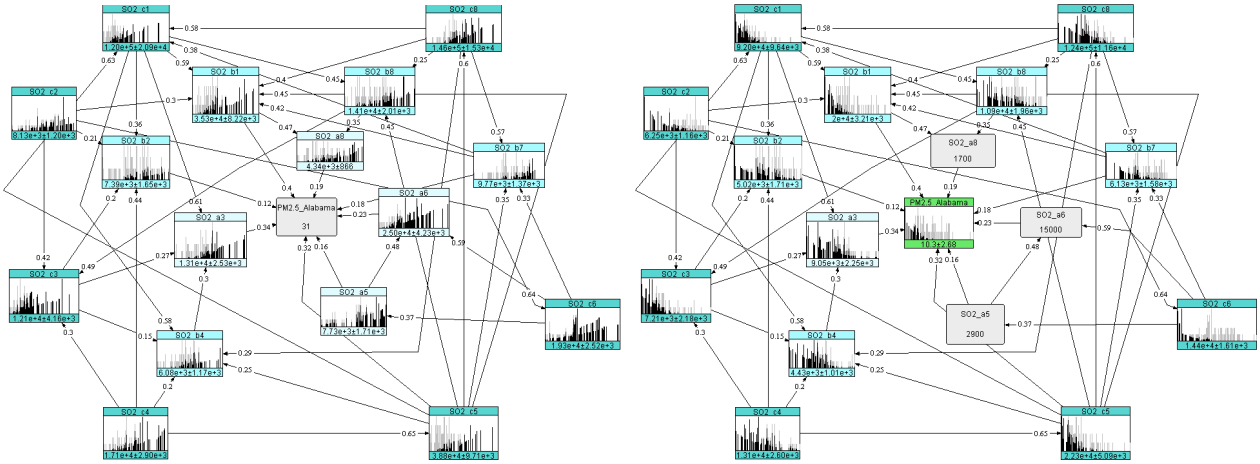
[3] All figures (except Figure 1a) are obtained with the software application UniNet.



(a) NPBBN on 17 variables. (b) Marginal distributions of the 17 variables.

Fig. 2. NPBBN for Alabama ambient $PM_{2.5}$.

the dependence structure present in the data set. The 17 variables will be nodes of a DAG. If the DAG has no arcs then DBBN = 1. In general, if the DAG is not saturated, then DBBN > DNR. Following the general procedure we start adding arcs between variables whose rank correlations are among the largest. By doing so, we decrease the value of DBBN. After a number of intermediate steps, we build a NPBBN with 35 arcs, most of which correspond to the highest rank correlations. Nevertheless, our interest is to quantify the relation between the $PM_{2.5}$ concentration and the rest of the variables, hence we also add arcs that carry information about their direct relationship. The resultant NPBBN has DBBN = 1.017E-03 and the 5% precentile of its distribution is 0.582E-04 which is larger than DNR. In consequence, we add 10 more arcs. The resultant structure is shown in Figure 2a. The determinant of the rank correlation matrix based on the new NPBBN is DBBN = 1.261E-04, its 90% central confidence band is $[0.059\text{E-}04, 1.296\text{E-}04]$ and DNR falls inside this interval. We conclude that this NPBBN is an adequate model of the saturated graph.



(a) Conditionalisation on high values of $PM_{2.5}$. (b) Conditionalisation on low values of SO_2 in ring "a".

Fig. 3. Diagnosis & prediction.

Once the NPBBN is learned from data, it can be used for inference. We can condition any set of variables on values of any other set of variables. Conditionalisation is performed on the transformed

variables, which are assumed to follow a joint normal distribution, hence any conditional distribution will also be normal with known mean and variance. Finding the conditional distribution of a corresponding original variable will just be a matter of transforming it back using the inverse distribution function of this variable and the standard normal distribution function [3].

Standard regression analysis also computes conditional distributions. For data sets like that encountered here, however, the NPBBN approach with the normal copula offers several advantages:

- We obtain the full conditional distribution, not just the mean and variance.
- We do not assume that the predicted variable has constant conditional variance, indeed the conditional distributions do not have constant variance.
- The emitters tend to be strongly correlated to each other and weakly correlated to the collectors, hence if we marginalize over a small set of emitters, we have many "missing covariates" with strong correlations to the included covariates. This will bias the estimates of the regression coefficients. The NPBBN method, in contrast, simply models a small set of variables, where other variables have been integrated out. There is no bias; the result of first marginalising then conditionalising is the same as first conditionalising then marginalising.
- The set of regressors may have individually weak correlations with the predicted variable, but may be collectively important. On small data sets, the confidence intervals for the regression coefficients may all contain zero and their collective importance would be missed.

We can use the learned NPBBN to answer questions such as: *Given measured ambient concentrations at one or more collectors, what can we infer about given emitters?*, or *Given emissions at a number of relevant plants, what is the effect on a given collector?*.

To illustrate, Figure 3a tries to answer a question of the first type, in the case of high $PM_{2.5}$ concentration. The differences between the emitters' conditional distributions (black), and the original ones (gray), caused by changing the concentration are striking. These differences are present for sources in all compass directions. In addition, emissions close to receptor tend to exhibit larger changes than emissions further away. Figure 3b answer a question of the second type. Conditionalisation was made on low values of SO_2 in ring "a". This results in a reduction of $PM_{2.5}$ concentration's mean from 15.4 to 10.3. The effects of different combinations of factors may be similarly investigated.

СПИСОК ЛИТЕРАТУРЫ

1. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufman Publishers, 1988.
2. Hanea A.M., Kurowicka D. Mixed Non-Parametric Continuous and Discrete Bayesian Belief Nets. *Advances in Mathematical Modeling for Reliability*, 2008, ISBN 978-1-58603-865-6, IOS Press.
3. Hanea A.M., Kurowicka D., Cooke, R.M. Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets. *Quality and Reliability Engineering International*, 2006, Vol. 22, No. 6, 613-729.
4. Hanea A.M., Kurowicka D., Cooke R.M., Ababei D.A. Mining and Visualising Ordinal Data with Non-Parametric Continuous BBNs. *Computational Statistics and Data Analysis*, 2008, doi:10.1016/j.csda.2008.09.032.
5. Pearson K. Mathematical contributions to the theory of evolution. *Biometric*, 1907, Series. VI.Series.