Unsupervised Learning - 24/25 - Exam Project

Alessio Zanga[1,2,⋆] and Fabio Stella[1]

[1] Models and Algorithms for Data and Text Mining Laboratory (MADLab),
Department of Informatics, Systems and Communication (DISCo),
University of Milano-Bicocca, Milan, Italy
[2] School of Medicine and Surgery,
University of Milano-Bicocca, Monza, Italy

1 Introduction

– You are given the `mehra-complete-1000.csv` dataset. The data has been generated from a Bayesian Network using the following `R` code:

```r
install.packages("bnlearn")
library(bnlearn)
load(url("https://www.bnlearn.com/bnrepository/mehra/mehra-complete.rda"))
set.seed(42)
data <- rbn(bn, n = 1000)
write.csv(data, "mehra-complete-1000.csv", row.names = FALSE)
```

– You can find the paper describing the dataset/model in the following pages.
– You are asked to perform the tasks listed in the next section.

2 Tasks

1. Read the attached paper and explore the dataset.
2. Apply at least three clustering techniques presented in this course.
3. Write a report on your findings:
   – What can we say about the `Type` variable?
   – How is `Type` affecting/affected by the other variables?
4. Submit the report in PDF format and the code as a Colab notebook.

3 Notes

– You are asked to carry out this project as a group project (up to 3 persons).
– You can use the e-learning forum to find or form a group.
– If you are a working student you are allowed to do it on your own.

⋆ Corresponding author: alessio.zanga@unimib.it

AGU PUBLICATIONS

Earth and Space Science

RESEARCH ARTICLE
10.1002/2017EA000326

Key Points:
- Bayesian Networks are a convenient type of models to investigate different sources of information at various temporal and spatial scales
- The methodology is scalable and can be applied from small to very large data sets
- For pollution and weather variables the model tests well in sample but also has good predictive power when tested out of sample

Correspondence to:
C. Vitolo,
claudia.vitolo@ecmwf.int

Citation:
Vitolo C., Scutari M., Ghalaieny M., Tucker A., & Russell A. (2018). Modeling air pollution, climate, and health data using Bayesian Networks: A case study of the English regions. *Earth and Space Science*, *5*, 76–88. https://doi.org/10.1002/2017EA000326

Received 4 AUG 2017
Accepted 22 JAN 2018
Accepted article online 26 JAN 2018
Published online 9 APR 2018

Modeling Air Pollution, Climate, and Health Data Using Bayesian Networks: A Case Study of the English Regions

Claudia Vitolo[1], Marco Scutari[2], Mohamed Ghalaieny[3], Allan Tucker[4], and Andrew Russell[3,5]

[1]Forecast Department, European Centre for Medium-range Weather Forecasts, Reading, UK, [2]Department of Statistics, Oxford University, Oxford, UK, [3]Institute of Environment, Health and Societies, Brunel University London, Uxbridge, UK, [4]Department of Computer Science, Brunel University London, Uxbridge, UK, [5]Committee on Climate Change, London, UK

Abstract The link between pollution and health is commonly explored by trying to identify the dominant cause of pollution and its most significant effect on health outcomes. The use of multivariate features to describe exposure is less explored because investigating a large domain of scenarios is theoretically (i.e., interpretation of results) and technically (i.e., computational effort) challenging. In this work we explore the use of Bayesian Networks with a multivariate approach to identify the probabilistic dependence structure of the environment-health nexus. This consists of environmental factors (topography and climate), exposure levels (concentration of outdoor air pollutants), and health outcomes (mortality rates). The information is collated with regard to a data-rich study area: the English regions (UK), which incorporate environmental types that are different in character from urban to rural. We implemented a reproducible workflow in the R programming language to collate environment-health data and analyze almost 50 millions of observations making use of a graphical model (Bayesian Network) and Big Data technologies. Results show that for pollution and weather variables the model tests well in sample but also has good predictive power when tested out of sample. This is facilitated by a training/testing split in the data along time and space dimension and suggests that the model generalizes well to new regions and time periods.

1. Introduction

There is an overwhelming body of evidence that environmental pollution, and air pollution in particular, is a significant threat to health worldwide. The World Meteorological Organization (World Health Organization, 2006) identifies six outdoor air pollutants for which there is strong and clear evidence of major impact on health: Ozone (O_3), Particulate Matters ($PM_{2.5}$ and PM_{10}), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Carbon Monoxide (CO). In Europe, health standards and objectives have been set by the European Commission with the introduction of the Commission of the European Communities (CEC) (2008). This Directive was made law in England through the Air Quality Standards Regulations in 2010 and specifies concentration limits for each major pollutant. However, the introduction of air pollutant concentration limits is only the first step toward the development of an air quality policy. Most importantly, in order to asses the potential human exposure and to estimate the long-term repercussions on population health, air quality scenarios need to be investigated. With this in mind, the paper here aims to investigate and test a novel, flexible framework for modeling the environment-health nexus.

According to Jerrett et al. (2005), population exposure to the above species can be modeled in various ways, from using simple proximity measures (Buzzelli & Jerrett, 2003; Ciccone et al., 1998), geospatial interpolation using kriging (Mulholland et al., 1998) and land use regression (Briggs et al., 1997; Ryan & LeMasters, 2007) models to more complex atmospheric dispersion (Caputo et al., 2003; Turner, 1979), integrated meteorological-emission (Gaines Wilson & Zawar-Reza, 2006), and hybrid models (Cauvin et al., 2001). De Hoogh et al. (2014) compared exposure patterns derived from land use regression models and dispersion models and found that results from these two models correlate well for NO_2, but the agreement is considerably lower for Particulate Matter (for both coarse and fine particles). This suggests that in order to generate accurate and unambiguous exposure patterns, more experiments are needed to integrate multiple model

strategies. Blangiardo et al. (2013) propose the use of a Bayesian approach to model spatiotemporal variability of air pollutants. There are a number of advantages in modeling scenarios using a formal Bayesian framework. First, expert opinion and literature results can be included in the analysis through the definition of a prior. Second, probabilities can be obtained from the posterior distribution, and, lastly, it is relatively easy to specify a hierarchical structure on data and parameters (useful to make predictions and working with missing data). Blangiardo et al. (2013), Brooks et al. (2011), and Robert and Casella (2013) seem to suggest that it is cumbersome to preserve spatial variability in order to improve the predictive power of statistical models. However, the authors only model a single pollutant at a time; this is a common limitation of many epidemiological studies. In recent years, instead, there has been increasing interest in the combined effect of multiple species. Therefore, the modeling work presented here will incorporate data from multiple species in an effort to understand the more holistic nature and impact of exposure.

In addition to pollution exposure, it is also important to consider environmental factors, some of which may have an exacerbating effect, some others a mitigating one. Extreme temperatures and ultraviolet radiation, for instance, can exacerbate preexisting health conditions such as cardiovascular and pulmonary diseases, as well as trigger new ones by aggravating exposure levels. Precipitation and wind can, instead, facilitate deposition and dispersion of pollutants. Basu and Samet (2002) reviewed a number of US studies on the relationship between elevated ambient temperature and mortality and concluded that the risk is higher for people with preexisting cardiovascular and respiratory diseases but that age and socioeconomic status can also be factors worth considering. Bell et al. (2004) analyzed the relation between Ozone and mortality, while temperature and other weather conditions were considered confounders. Estimating the interaction between temperature and Ozone in light of observed health outcomes is not trivial, as these are highly correlated variables. Jhun et al. (2014) assessed Ozone-related mortality risk considering, on the one hand, the temperature as confounder and effect modifier and, on the other hand, air conditioning prevalence in 97 cities in the United States. They found a statistically significant increase in Ozone mortality risk during high temperature days and also that air conditioning seemed to have a mitigating effect. As such, the present study includes certain environmental variables beyond air pollution in an effort to capture more of the factors that combine to influence mortality.

Zheng et al. (2013) investigated, for Beijing and Shangai, the interaction of SO_2, NO_2 as well as fine and coarse Particulate Matter ($PM_{2.5}$ and PM_{10}) with hourly meteorological data and found the following to be the most relevant: high wind speed is associated with lower concentration of PM_{10}; and high humidity causes high concentration of PM_{10}. In general, air quality is good in two cases: high temperature and low humidity; and high pressure and low temperature. De Sario et al. (2013) showed that urban European cities are also highly vulnerable areas and highlighted how extreme weather patterns/events and changes in the concentration of pollutants/aeroallergens have synergistic effects: increasingly high temperature and sunnier days are associated with an increase in O_3 and SO_2, while high precipitation facilitates the sinking of Particulate Matter but could also increase SO_2 (because of the additional water vapor). These factors increase allergic reactions, decline lung function, cause lung cancer, and even premature death.

IMPORTANTE!

Modeling the effects of ambient air pollution on health taking all the relevant variables into account is rather challenging from a theoretical point of view, as the environment-health nexus is expected to be characterized by a rather complex dependence structures. Computational challenges are also apparent, since analyzing large databases requires long processing times and can only be handled with an adequate computer infrastructure in place. When resources are limited and/or research is driven mainly by openly available data, there is a need for models that can intrinsically identify dependencies based on the variability of the different features and can ingest expert knowledge to boost predictability. In this context, graphical models such as Bayesian Networks (BNs) seem to be a perfect fit. These models are increasingly being used in computer science (Hu et al., 2013) and business analytics problems (Duan & Xiong, 2015) as well as medicine (Wilson et al., 2015), genetics (Scutari et al., 2014), and epidemiology (Lappenschaar et al., 2013) while relatively less explored in environmental sciences (Aguilera et al., 2011). BNs can be inferred from data, their construction involves identifying the conditional independence structures among variables (or their joint probability distribution), and are schematized as a Directed Acyclic Graph (DAG) in which features are represented as nodes and dependencies as edges. By definition, a DAG cannot contain cycles and an edge can only have one direction. The advantage of using such models is threefold: (1) they limit the number of possible dependencies to analyze during the structure learning task making the network easier to inspect visually; (2) speed up computations (the fewer dependencies, the less computational time is required); and finally (3) they allow the introduction

of expert knowledge. The latter can be done in a number of ways, including constraints on the parameters of the models (Friedman & Goldszmit, 1998) and by whitelisting or blacklisting specific edges in the network (Scutari, 2010); in this work we adopted the latter approach for computational reasons.The whitelist declares which edges are forced to be present in the DAG, while the blacklist declares edges that are excluded from the DAG. Whitelists and blacklists are inherently subjective as they depend on the expert/modeler's knowledge of the phenomenon which can vary based on the location, scale of the analysis, and data availability constraints. Therefore, they cannot be generalized but need to be formulated on a project-by-project basis. Given their successful use in Big Data analyses, BNs will be used in this paper in an attempt to model aspects of the environment-health system.

1.1. Paper Aim and Outline

The main goal and novelty of this work consists of investigating whether BNs can be used with large volumes of heterogeneous data (in terms of spatiotemporal scale and data types) and still able to identify, interpret, and predict the dependence structure between these predictors and health outcomes (mortality). As described in section 1, this has significant value in understanding the complex links between environmental factors and health outcomes as well as being used in the evidence base to inform policy interventions. To the best knowledge of the authors, the variety and volume of information taken into account has not been previously analyzed for the English regions and constitutes an additional novelty of this work. The remainder of the paper is organized as follows: in section 2 we describe the case study and data availability, as well as how we suggest to assemble the database of available information, handle missing values and build the BN for both continuous and categorical variables. The results of the structure learning process, inference, and predictions are discussed in section 3, while the overall results are discussed in section 4 and main conclusions and future works are summarized in section 5.

2. Data and Methods

According to the guidelines suggested by Marcot et al. (2006) and Kalisch et al. (2012), we built and revised the BN through a sequence of steps: (1) feature identification, (2) structure learning, and (3) validation.

2.1. Feature Identification

We take into account the air quality monitoring stations in England (UK) as location of interest and extract the weather, geography, and health data at these locations, from 1981 to 2014. The recorded features are summarized in Table 1. For each variable, the column named "Type" shows whether the variable is continuous (C) or discrete (D). The table also contains the names of variables as used by the model, which is helpful for reading the network and interpret the results. More details are given in the following subsections.

2.1.1. Pollution Data

We identified relevant features for air pollution modeling reviewing the literature in the field. In particular, the exposure is calculated at the location of air pollution monitoring stations, taking into account temporal factors (year, season, month, day, and time of measurements), as well as the environmental factors in terms of weather variables (2 m temperature, 10 m wind speed and direction, total precipitation, boundary layer height, and surface net solar radiation) and pollutant species (Ozone, fine and coarse Particulate Matter, Sulfur Dioxide, Nitrogen Dioxide, and Carbon Monoxide). The UK Air Information Resource service hosted by the Department for Environment, Food, and Rural Affairs (DEFRA) includes thousands of air quality monitoring stations. Many, however, use obsolete sensors and/or have been dismissed. The most reliable measurements in England are available from a network of 162 stations, which record hourly measurements. These stations were identified using the *rdefra* R package (Vitolo et al., 2017; Vitolo et al., 2017a; Vitolo, Russell, & Tucker, 2016). The geographical distribution of data points is shown in Figure 1. The *openair* R package (Carslaw & Ropkins, 2012, 2016) was used to import the related hourly time series. The temporal coverage is highly variable, with a minimum of 4 months and a maximum of 35 years (average: 12 years). Spatially, the stations are evenly distributed across regions but concentrated in urban areas within each region (see Figure 2). As a result some variables have zero or near-zero variance in the BN initialization stage, in which only complete observations are taken into account. In particular, complete observations are only recorded for the period 1998–2000 in the urban area of the Greater London Authority (Environment Type: Background Urban Traffic). Since such variables are non-informative under the distributional assumptions used for structure learning below (Kuhn & Johnson, 2013), we disregard them (region, zone, type, and year) when running the EM algorithm. Individual stations usually monitor only a subset of pollutants. For instance, we obtained O_3 data from 95 stations for an average

Table 1
Data Summary Table

Category	Feature	Name	Type	Time step	Unit	Temporal coverage	Spatial coverage
Health outcomes	Mortality rates	CVD60	C	1 day	—	35 years	English regions
Air pollution	Ozone	O_3	C	1 h	$\mu g/m^3$	14 years	95 stations
	Particulate Matter with $d < 2.5\mu m$	$PM_{2.5}$	C	1 h	$\mu g/m^3$	6 years	64 stations
	Particulate Matter with $d < 10\mu m$	PM_{10}	C	1 h	$\mu g/m^3$	11 years	83 stations
	Sulfur dioxide	SO_2	C	1 h	$\mu g/m^3$	12 years	85 stations
	Nitrogen dioxide	NO_2	C	1 h	$\mu g/m^3$	12 years	146 stations
	Carbon monoxide	CO	C	1 h	$\mu g/m^3$	11 years	80 stations
Weather	Wind speed	WS	C	3 h	m/s	35 years	162 stations
	Wind direction	WD	C	3 h	Degrees	35 years	162 stations
	Temperature	T2 M	C	3 h	K	35 years	162 stations
	Total precipitation	TP	C	3 h	mm	35 years	162 stations
	Boundary layer height	BLH	C	3 h	m	35 years	162 stations
	Surface solar net radiation	SSR	C	3 h	$W/m^2 s$	35 years	162 stations
Time of	Year	YEAR	D	—	—	—	—
pollution	Season	SEA	D	—	—	—	—
measurement	Month	MON	D	—	—	—	—
	Day	DAY	D	—	—	—	—
	Hour	HOUR	D	—	—	—	—
Geography	Longitude	LON	C	—	Degrees	—	162 stations
	Latitude	LAT	C	—	Degrees	—	162 stations
	Altitude	ALT	C	—	mAOD	—	162 stations
	Region	REG	D	—	—	—	—
	Zone	ZONE	D	—	—	—	—
	Environmental type	TYPE	D	—	—	—	—

Note. Features can be of two types: continuous (C) or discrete/categorical (D).

of 14 years, while $PM_{2.5}$ was measured in only 64 stations for an average of 6 years. From the same source, we also obtained the exact location (latitude, longitude, and altitude) and environmental type of each station. Missing altitude values were imputed by point inspection of the Ordnance Survey Terrain 50, a Digital Terrain Model characterized by a spatial resolution of 50 m.

2.1.2. Weather Data

Weather information was obtained from ERA-interim (Dee et al., 2011), a reanalysis data product developed by the European Centre for Medium-range Weather Forecasts. This is available in a gridded format with a spatial resolution of about 80 km and a temporal resolution of up to 3 h. ERA-interim data were retrieved via the ECMWF-MARS web service, imported using the *ncdf4* R package (Pierce, 2015), and the climate variables at each station were extracted by point inspection using the *raster* R package (Hijmans, 2016). The script used to generate the results is based on the *kehra* R package (Vitolo, Tucker, & Russell, 2016).

2.1.3. Health Data

Health outcomes are described in terms of mortality rates. These are obtained from mortality counts per thousand individuals split by region, age, and day of occurrence. Data are provided by the Office for National Statistics and cumulated within each region of England (Office of the National Statistics, 2016a, 2016b). The data on mortality were filtered for the over 60 age demographic in order to focus on a vulnerable population group. The choice of this age bracket further necessitated the aggregation of data on the larger scale of the English regions because the use of any smaller geographical scale could have resulted in the identification of individuals, which is not permitted with this data set.

In terms of causes of death, we only take into consideration cardiovascular-pulmonary diseases (CVD) and cancer for which the International Classification for Diseases (ICD) codes are as follows:

1. ICD10 codes for CVD: I00-I99, J00-J99 (period 2001 – 2014);
2. ICD9 codes for CVD: 390-459, 460-519 (period pre-2001);
3. ICD10 codes for cancer: C30-39, C45;
4. ICD10 codes for cancer: 160-165.

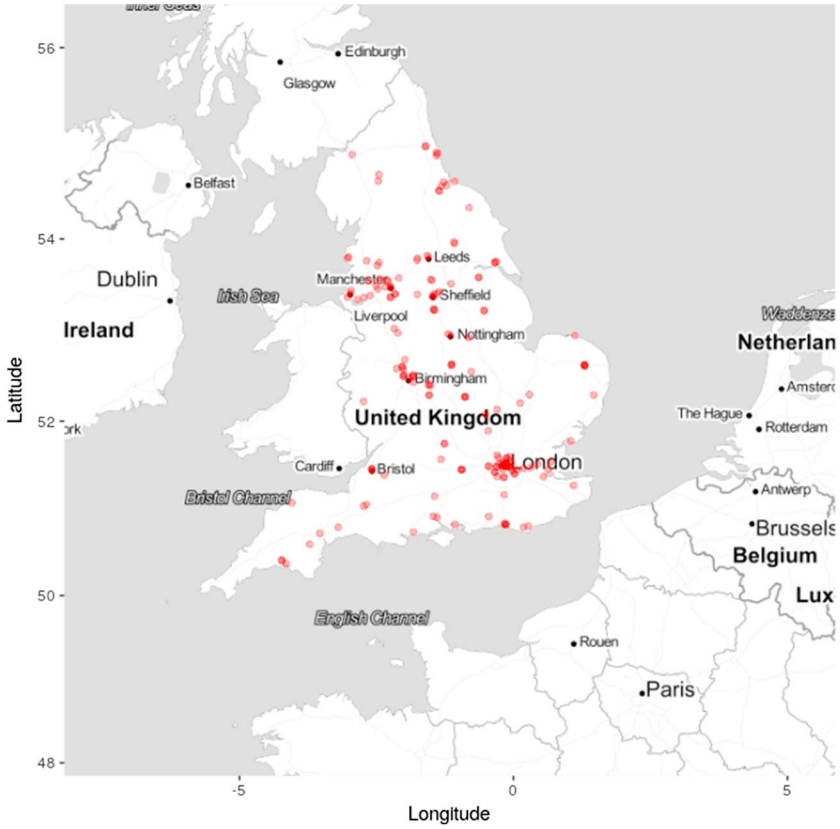

Figure 1. Air quality monitoring stations in England.

Lastly, mortality rates are calculated by dividing the mortality counts by yearly regional population estimates obtained from the MYEDE data set (Office for National Statistics, 2015). The population estimates are considered constant over the year and the mortality rates constant over each region.

2.2. Structure Learning

A BN is built to identify the dependence structure among exposure and outcome variables. To learn the model structure, we used the Hill Climbing (HC) algorithm (Russell & Norvig, 2016) as implemented in the *bnlearn* package (Scutari, 2010). HC performs a greedy search starting from an initial DAG, which in our case contains no arcs, and evaluates different DAGs by iteratively adding/removing/reversing each possible arc and then keeping the DAG that fits the data best in each step. We estimate goodness of fit using the Bayesian Information Criterion (Schwarz, 1978), which approximates the posterior probability of the DAG.

Since we decided to include both discrete and continuous features in the analysis, we assume that the BN follows the Conditional Linear Gaussian distributional assumptions (Lauritzen & Wermuth, 1989). In particular, we assume that discrete features are categorical (i.e., their values are not ordered) and that continuous features can depend on discrete variables but not vice versa. The distribution of continuous features conditional on the respective parents is assumed to take the form of a set of classic linear regression models (one for each combination of the possible values of the discrete parents) in which the continuous parents take the role of explanatory variables.

In order to avoid inadvertently introducing bias in the BN, we decided not to declare any whitelist but only a blacklist marking some edges as unrealistic. In particular, topographic variables (latitude, longitude, and altitude) can influence weather and pollutants but not vice versa; pollutants can influence weather and mortality but not vice versa; and mortality rates cannot influence any of the other variables.

2.3. Imputation and Learning

The assembled data set consists of almost 50 million records with 24 features. We split it into a training and testing set, comprising the records in 1981–2005 (74%) and in 2006–2014 (26%), respectively. Training and testing data sets are publicly available (Vitolo et al., 2017b).

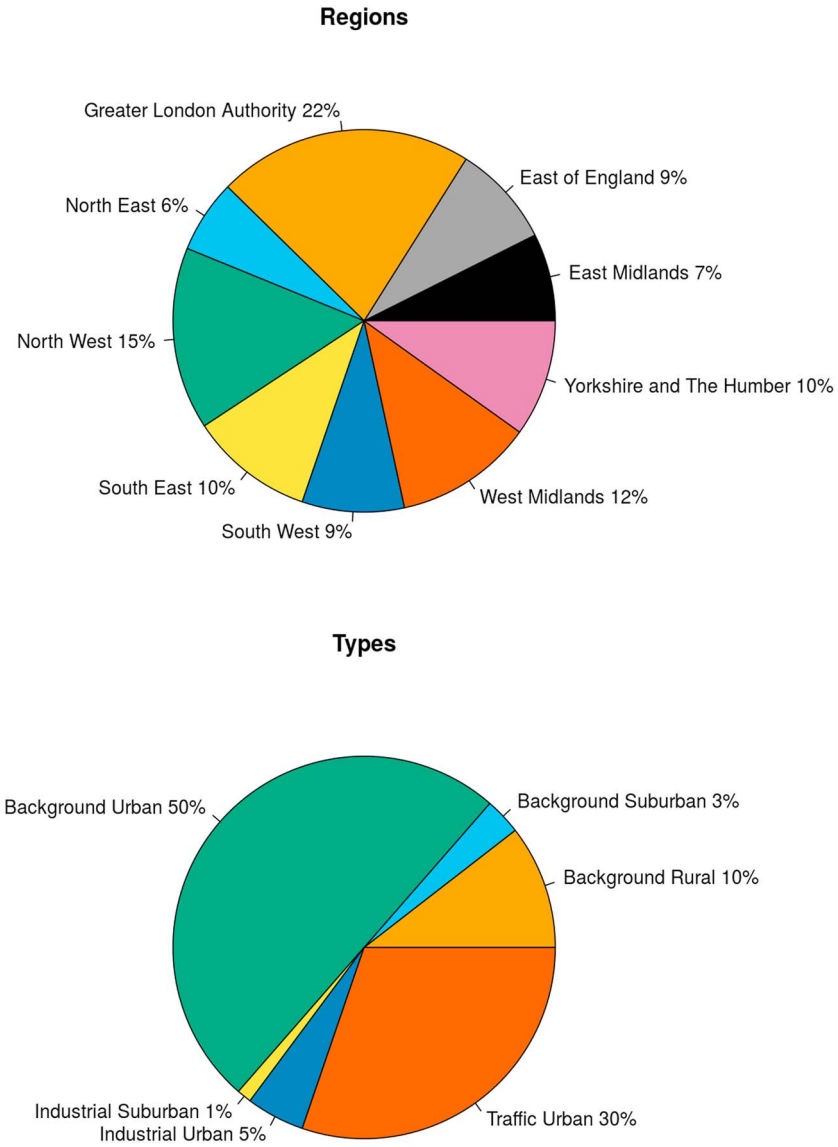

Figure 2. Distribution of pollution measurements, regions are overall homogeneously represented (top chart), while the environmental type is biased toward urban areas (bottom chart).

The gaps between each pair of consecutive weather observations in the training set are filled in using linear interpolation. The missing values in pollution measurements, on the other hand, are incorporated in model estimation using the Structural Expectation-Maximization algorithm (Friedman, 1997):

1. Define any blacklist/whitelist.
2. Initialize using the empty graph and complete observations.
3. Fit the parameters for the empty structure using their maximum likelihood estimates.
4. Repeat the following until convergence:
    a. Expectation step: replace missing values with their posterior expectations conditional on the observed values, using the predict() function.
    b. Maximization step: learn the model that maximizes the score with the current data, using the HC algorithm to learn the DAG and the bn.fit() function to learn the parameters of the related DAG.

This is initialized using an empty graph and the blacklist. Then missing values in each observation are imputed with their maximum a posteriori estimates from the variables that are observed, and a new graph is learned from the now complete data. These two steps, imputation and learning, are repeated until the learned DAG

Table 2
Comparison of Successive Iterations of the EM Algorithm, Where Every Iteration i is Compared With Iteration i+1

	Common arcs	Arcs added			Arcs removed		
Iterations	#	#	from	to	#	from	to
1–2	68	1	BLH	O_3	0	—	—
2–3	69	0	—	—	0	—	—
3–4	69	0	—	—	0	—	—
4–5	68	1	*CO*	NO_2	1	NO_2	*CO*
5–6	68	1	NO_2	*CO*	1	*CO*	NO_2
6–7	68	2	SO_2	SSR	1	NO_2	*CO*
			CO	NO_2		—	—
7–8	70	1	BLH	NO_2	0	—	—

Note. The number of common arcs are those arcs that are unchanged from iteration i to iteration i + 1 . The number of arcs added corresponds to the number of arcs in iteration i + 1 not present in iteration i. The number of arcs removed corresponds to the number of arcs not in iteration i + 1 but present in iteration i.

and the imputed data are stable (i.e., they do not change significantly). Final BN model and related DAG, which are the same for all the English regions, are publicly available (Vitolo et al., 2017b).

2.4. Validation

The training data set is used to generate the graph model. After learning the structure and parameters of the BN, we assess its accuracy through the analysis of residuals and validate it by comparing predicted variables under unobserved conditions provided by the testing data set.

2.5. A Note on Managing Big Data

The method described above produces data sets whose size depends on the number of monitoring stations and temporal coverage of the network. The more data-rich the area, the larger the data set becomes. This has a strong impact on the performance of the analysis, which can take a long time (if it is possible at all) for a data set made of several millions of records and tens of features, using an average desktop machine.

Determining the most appropriate technologies to employ, both in terms of hardware and software, is crucial in this respect. We decided to speed-up the learning process by distributing the calculations over multiple cores via the *parallel* R package (R Core Team, 2016) and relying on a high-end server designed for high-performance computing, see Appendix B. Without going into much detail on the parallelization, which is beyond the scope of this paper, horizontal scaling was essential to build the database and run the Structural EM algorithm, while vertical scaling was used for both exploratory analysis and verification of results. We developed this analysis pipeline in the R programming language because of the availability of libraries implementing most of the required algorithms. These libraries have been thoroughly tested and, in most cases, are considered the reference implementations of the methods they implement.

3. Results

Imposing a wall time of 2 months, the EM algorithm iterated eight times. The calculation did not fully converge; however, for every successive iteration the structures showed a maximum of three different arcs only, while at least 68 arcs were in common (see Table 2).

3.1. Network Structure

The DAG obtained from the last iteration (Figure 3) was used for the subsequent analysis. This structure clearly detects the hierarchical structure of the different scales of observation, with the mortality over the age of 60 (CVD60, observed at regional scale) related to the geographical location (Region) and variable with time (year, season, and month). Within each region, the proximity to urban areas (encoded in the TYPE and ZONE variables) and the time of the year affect the concentration of pollutants. These, in turn, influence the weather.

3.2. Analysis of Residuals and Network Validation

The accuracy of the model is assessed by analyzing the residuals between the observed variables and those predicted by the model. In Table 3, the root mean square error (RMSE) is used to summarize the average

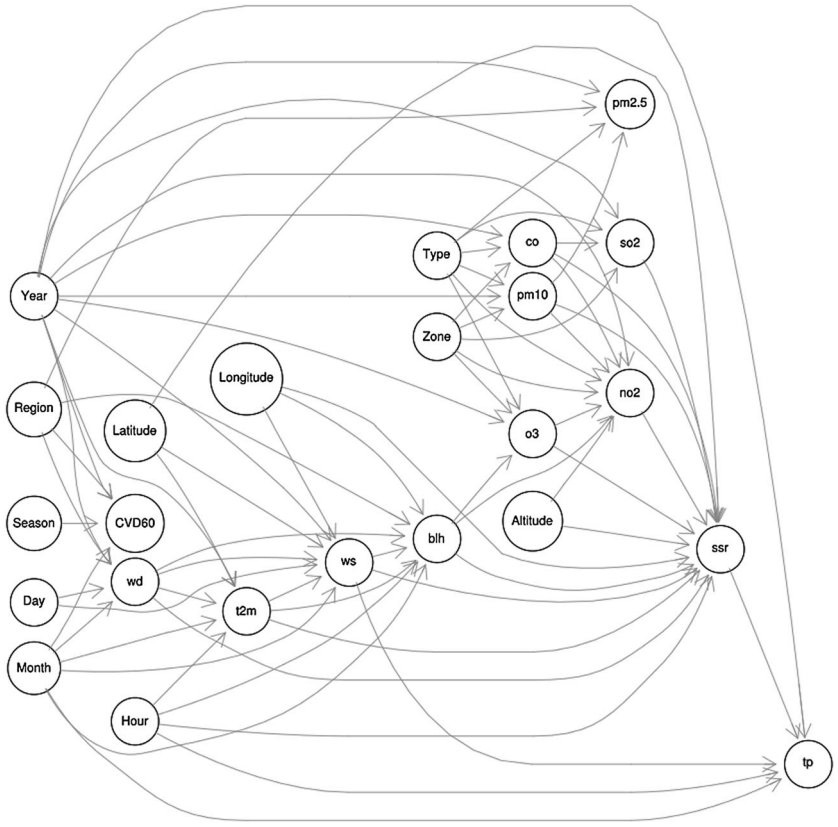

Figure 3. DAG describing dependence structure.

Table 3
RMSE for the Continuous Features in the Training and Testing Data Sets

Training		Testing	
Feature	Normalized RMSE	Feature	Normalized RMSE
O_3	0.05	O_3	0.12
$PM_{2.5}$	0.02	$PM_{2.5}$	0.03
PM_{10}	0.00	PM_{10}	0.03
SO_2	0.01	SO_2	0.03
NO_2	0.01	NO_2	0.08
CO	0.03	CO	0.06
Longitude	0.25	Longitude	0.40
Latitude	0.21	Latitude	0.49
Altitude	0.10	Altitude	0.94
Wind speed	0.04	Wind speed	0.19
Wind direction	0.14	Wind direction	0.34
Temperature	0.05	Temperature	0.09
Rainfall	0.03	Rainfall	0.06
Boundary layer height	0.07	Boundary layer height	0.10
Solar radiation	0.11	Solar radiation	0.13
Mortality rates	0.04	Mortality rates	0.27

deviation of the estimates from the actual values. The RMSE here is normalized to make it comparable across variables. As the training and testing data sets were split based on the Year, we disregarded this information when predicting using the testing data set.

The table shows that RMSE for pollution and weather variables in the training and testing sets are generally very similar. This suggests that the model tests well in sample but also has good predictive power when tested out of sample. Clear exceptions are the geographical variables longitude, latitude, and altitude for which errors increase from 0.25 – 0.21 – 0.10 to 0.40 – 0.49 – 0.94, respectively. The error associated to the mortality rates also increases considerably, from 0.04 to 0.27. This is most likely an overfitting issue, whereby the model seems to capture the underlying relationship among features as well as noise and spurious patterns in the training set. This could be a sign that the model is excessively complex and/or the database was not split in the ideal proportions (74% training set and 26% test set).

4. Discussion

This work presents a data-driven application in which we use BNs to model the statistical dependencies between environmental parameters, air pollution variables, and health outcomes. The input data are highly heterogeneous both in space and time. Although the outcome variable is associated to the English regions, we decided not to aggregate the environmental data at regional level because this would have caused a loss of information. We used, instead, the air quality monitoring stations as location of interest and extracted the weather, geography, and health data at these locations. This is based on the assumption that the dependence structure should be able to show different behaviors from one region to another, even though data are collated for a limited number of points.

As the air pollution data set was not complete, an expectation-maximization algorithm was used to make use of partial observations with missing values as part of the bnlearn implementation which generated the DAG. The process iterated eight times under an imposed time limit and was deemed to have effectively converged on the basis of the definition in Hand et al. (2001) which bases convergence on the lack of "appreciable difference between the final few iterations" of the process, and as Table 2 demonstrates, the number of different arcs in the final few iterations is less than two. The successful application of an EM algorithm is of great potential to this type of environmental health analysis as the availability of input data, particularly, on air pollution is often sparse. Therefore, with the aid of an EM algorithm, BN can be used to predict health outcomes with a degree of confidence despite the lack of total coverage for air pollution and other data. The ability of a BN model to make public health predictions with incomplete data is a major advantage; yet it should be noted that there is scope for further enhancement by either increased in situ air pollution measurements (e.g., using low-cost sensor technology) or by using satellite measurements of air quality and, therefore, improving the predictive power of the BN model.

A comparison between the DAG structure and known interrelations in atmospheric chemistry, meteorology, and health is a logical means of probing how well the BN model represents the real-world data it describes. The key outcome to be predicted is CVD60, and this is shown to be influenced by air pollution and meteorological variables as is well established in the literature (World Health Organization, 2006). However, the effect on CVD60 by air pollutants (O_3, NO_2, SO_2, PM_{10}, $PM_{2.5}$, and *CO*) and weather variables (*WS*, *SSR*, *BLH*, *T2M*, *WD*, and *TP*) appears to be mediated by the variables Year, Region, and Month. This is understandable as the pollution and weather variables naturally exhibit temporal and spatial variations. Nonetheless, it would be interesting to investigate the strength of direct arcs linking CVD60 and the pollution and weather variables by removing the intermediate variables in future work.

The model represents known processes in atmospheric chemistry with a good degree of accuracy. The arcs in the DAG connecting NO_2 (Nitrogen Dioxide), O_3 (Ozone), and *SSR* (surface solar radiation) indicate that the model captures the importance of photochemistry of Nitrogen Dioxide for the production of Ozone at ground level (Finlayson-Pitts & Pitts, 2000).

The direct arcs in the DAG from the variable *CO* (Carbon Monoxide) to SO_2 (Sulfur Dioxide) and NO_2 (Nitrogen Dioxide) are as expected for these primary pollutants that result directly from combustion (Finlayson-Pitts & Pitts, 2000). It would be interesting to explore whether the almost 95% reduction in sulfur emissions from coal burning during the period covered by the data set (NAEI, undated) would weaken the strength of that arc between *CO* and SO_2.

Although the DAG does not show any direct arcs from PM_{10} (Particulate Matter under 10 μm) and $PM_{2.5}$ (Particulate Matter under 2.5 μm) to *CO*, this relationship is mediated by NO_2 suggesting that the model may be showing primary aerosol production mediated by NO_2 as well as capturing secondary aerosol production (Finlayson-Pitts & Pitts, 2000) (e.g., in the form of ammonium nitrate aerosols), in addition to the influence of Nitrogen Dioxide on secondary organic aerosol production (Kroll et al., 2006). While these shortcomings do not have bearing on the predictive power of the model for health outcomes, which was assessed by an analysis of the differences between RMSE in the training and testing sets, they would benefit from further examination in future work to assess the overall predictive capacity of the model.

We also note that the time-related variables could have been encoded in more useable form. Year was mistakenly encoded as a categorical variable, thus introducing a limitation on the production of temporal trends. Furthermore, the mistaken treatment of Day, Month, and Hour as individual, categorical variables may have hampered the exploration of the results when in fact, it would have been more informative to encode all the time variables as a single variable using the Coordinated Universal Time format.

Results show that for pollution and weather variables the model tests well in sample (using the training set) but also has good predictive power when tested out of sample (using the testing set). The errors arising from the test data sets are, as expected, higher than errors arising from train data. As the difference is often substantial, the issue is probably due to overfitting. This could be addressed in future works by splitting the data so that 70% is used for training the model and the remaining 30% for testing.

To the best knowledge of the authors, a statistical analysis of the variety and volume of information taken into account has not been previously attempted, at least for the English regions, and constitutes the main novelty of this work. The closest attempt to investigate the effect of air pollution on mortality rates was made in a recent data science competition (Kaggle inClass competition on "Predict impact of air quality on mortality rates": https://www.kaggle.com/c/predict-impact-of-air-quality-on-death-rates), but the data set consisted of fewer features and spanned a shorter time range (2007–2014), also the air quality information was generated by averaging gridded data from the Copernicus Atmosphere Monitoring Service rather than looking at point-based information from the UK-Air Information Resource, as done in this study. The winning modeling approach was based on the eXtreme Gradient Boosting algorithm and resulted into an RMSE of 0.29 on the testing data set which is only slightly worse than the 0.27 scored by the BN in this work. We think, however, that the BN approach is more flexible (using a mixture of point-based and gridded data) and has the potential to further improve if fed by more detailed mortality rates. This model is also a valuable scenario exploration tool that can be used to support decision and policy makers. It can be used, for instance, to assess changes in mortality due to more extreme weather condition, concentration of pollutants, or a combination of the two by simulating the conditions the model would expect to observe in those adverse scenarios.

5. Conclusions

This work set out to determine whether a BN graphical probabilistic model could be used to identify and predict dependencies between variables that predict exposure to pollutants and population health outcomes. The analysis of residuals confirmed that BNs are a promising method for the use of multivariate environmental and air pollution data to predict health problems. Despite a few shortcomings, discussed above, the DAG structure accurately represents known linkages between the environment and health and also known processes within the environment.

We can conclude that this application of BN graphical predictive modeling offers great potential when exploring the environment-public health nexus as the model was able to process and analyze the multitude of variables involved, in addition to utilizing an EM algorithm to compensate for missing values. The ability to effectively deal with missing measurements would be of particular importance if the model were to be applied to environmental problems where the data availability is even more lacking and this work could be an extremely useful tool to provide statistically sound evidence to aid in public health policy decision making.

Future work to build upon this research includes the implementation of simulated scenario modeling to assess the effects of environmental change on CVD60 and to test the accuracy and effectiveness of model predictions of the environmental and weather variables. A series of comparisons between this machine learned BN model and models constructed from expert knowledge in addition to a model that is a hybrid of a machine learning and expert knowledge. Finally, this BN model was trained on data with good spatial coverage and relatively

high temporal resolution. It would be informative to check whether the same model could be used to predict CVD60 from air pollution and weather data sets in areas where data quality is not as good.

Appendix A: Hardware and Software Specifications

Hardware:

1. System: PowerEdge R815 (x3)
2. Processor: AMD Opteron(tm) Processor 6378 (64 cores)
3. Memory: 256GiB System Memory
4. Disk: 4998GB PERC H700

Software:

1. Platform: x86_64-pc-linux-gnu (64-bit)
2. Operating System: Ubuntu 14.04.4 LTS
3. R version 3.3.0 (2016-05-03)

Acronyms

BN	Bayesian Network
CVD	Cardiovascular-pulmonary diseases
DAG	Direct acyclic graph
DEFRA	Department for Environment Food and Rural Affairs (UK)
ECMWF	European Centre for Medium-Range Weather Forecasts
EM	Expectation Maximization
HC	Hill Climbing
ICD	International Classification of Diseases
RMSE	Root Mean Square Error

Acknowledgments
This research was carried out when Claudia Vitolo was working at Brunel University London as part of the project "A multidimensional environment-health risk analysis system for Kazakhstan," supported by the British Council Institutional Links grant 172614334. We thank the UK Office for National Statistics for providing the health data and make it publicly available afterward under the Freedom of Information Act.
We would also like to show our gratitude to David C. Carslaw (University of York) who provided great support to identify air pollution data availability and Anna Esposito (University of Naples, Italy) for assistance with the selection of the most appropriate ICD codes. Input data set and outputs of this work are available under the name "Multidimensional Environment-Health Risk Analysis (MEHRA) data and model for the English regions" hosted by the Zenodo public repository (Vitolo et al., 2017b).

References

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software*, *26*(12), 1376–1388. https://doi.org/10.1016/j.envsoft.2011.06.004

Basu, R., & Samet, J. M. (2002). Relation between elevated ambient temperature and mortality: A review of the epidemiologic evidence. *Epidemiologic reviews*, *24*(2), 190–202. https://doi.org/10.1093/epirev/mxf007

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., & Dominici, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987-2000. *Jama*, *292*(19), 2372–2378. https://doi.org/10.1001/jama.292.19.2372

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, *7*, 39–55. https://doi.org/10.1016/j.sste.2013.07.003

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., … Van Der Veen, A. (1997). Mapping urban air pollution using GIS: A regression-based approach. *International Journal of Geographical Information Science*, *11*(7), 699–718. https://doi.org/10.1080/136588197242158

Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Buzzelli, M., & Jerrett, M. (2003). Comparing proximity measures of exposure to geostatistical estimates in environmental justice research. *Global Environmental Change Part B: Environmental Hazards*, *5*(1), 13–21. https://doi.org/10.1016/j.hazards.2003.11.001

Caputo, M., Giménez, M., & Schlamp, M. (2003). Intercomparison of atmospheric dispersion models. *Atmospheric Environment*, *37*(18), 2435–2449. https://doi.org/10.1016/S1352-2310(03)00201-2

Carslaw, D. C., & Ropkins, K. (2012). Openair — An R package for air quality data analysis. *Environmental Modelling & Software*, *27–28*, 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008

Carslaw, D. C., & Ropkins, K. (2016). Openair: Open-source tools for the analysis of air pollution data. R package version 1.8–6. Retrieved from https://CRAN.R-project.org/package=openair

Cauvin, S., Moullec, Y. L., Bremont, F., Momas, I., Balducci, F., Ciognard, F., … Zmirou, D. (2001). Relationships between nitrogen dioxide personal exposure and ambient air monitoring measurements among children in three French metropolitan areas: VESTA study. *Archives of Environmental Health: An International Journal*, *56*(4), 336–341. https://doi.org/10.1080/00039890109604465

Ciccone, G., Forastiere, F., Agabiti, N., Biggeri, A., Bisanti, L., Chellini, E., … Viegi, G. (1998). Road traffic and adverse respiratory effects in children. SIDRIA Collaborative Group. *Occupational and Environmental Medicine*, *55*(11), 771–778. https://doi.org/10.1136/oem.55.11.771

Commission of the European Communities (CEC) (2008). CEC (Commission of the European Communities) Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*, *L152*, 1–44.

De Hoogh, K., Korek, M., Vienneau, D., Keuken, M., Kukkonen, J., Nieuwenhuijsen, M. J., … Bellander, T. (2014). Comparing land use regression and dispersion modelling to assess residential exposure to ambient air pollution for epidemiological studies. *Environment international*, *73*, 382–392. https://doi.org/10.1016/j.envint.2014.08.011

De Sario, M., Katsouyanni, K., & Michelozzi, P. (2013). Climate change, extreme weather events, air pollution and respiratory health in Europe. *European Respiratory Journal*, *42*(3), 826–843. https://doi.org/10.1183/09031936.00074712

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., … Vitart, F. (2011). The ERA-Interim Reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*, 553–597. https://doi.org/10.1002/qj.828

Duan, L., & Xiong, Y. (2015). Big data analytics and business analytics. *Journal of Management Analytics*, *2*, 1–21. https://doi.org/10.1080/23270012.2015.1020891

Finlayson-Pitts, B. J., & Pitts, J. N. (2000). *Chemistry of the upper and lower atmosphere: Theory, experiments and applications*. San Diego, CA: Academic Press.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In D. H. Fisher (Ed.), *Proceedings of the 14th International Conference on Machine Learning (ICML '97)* (pp. 125–133). San Francisco, CA: Morgan Kaufmann Inc.

Friedman, N, & Goldszmit, M. (1998). Learning Bayesian networks with local structure. In Jordan M. I. (Ed.), *Learning in Graphical Models*, NATO ASI Series (Series D: Behavoioural and Social Sciences). Dordrecht: Springer.

Gaines Wilson, J., & Zawar-Reza, P. (2006). Intraurban-scale dispersion modelling of particulate matter concentrations: Applications for exposure estimates in cohort studies. *Atmospheric Environment*, *40*(6), 1053–1063. https://doi.org/10.1016/j.atmosenv.2005.11.026

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: The MIT Press.

Hijmans, R. J. (2016). Raster: Geographic data analysis and modeling. R package version 2.5-8. Retrieved from https://CRAN.R-project.org/package=raster

Hu, Y., Zhang, X., Ngai, E. W. T., Cai, R., & Liu, M. (2013). Software project risk analysis using Bayesian networks with causality constraints. *Decision Support Systems*, *56*, 439–449. https://doi.org/10.1016/j.dss.2012.11.001

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., … Giovis, C. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, *15*(2), 185–204. https://doi.org/10.1038/sj.jea.7500388

Jhun, I., Fann, N., Zanobetti, A., & Hubbell, B. (2014). Effect modification of ozone-related mortality risks by temperature in 97 US cities. *Environment international*, *73*, 128–134. https://doi.org/10.1016/j.envint.2014.07.009

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, *47*(11), 1–26.

Kroll, J. H., Ng, N. L., Murphy, S. M., Flagan, R. C., & Seinfeld, J. H. (2006). Secondary organic aerosol formation from isoprene photooxidation. *Environmental Science & Technology*, *40*(6), 1869–1877. https://doi.org/10.1021/es0524301

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lappenschaar, M., Hommersom, A., Lucas, P. J. F., Lagro, J., Visscher, S., Korevaar, J. C., & Schellevis, F. G. (2013). Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity. *Journal of Clinical Epidemiology*, *66*(12), 1405–1416. https://doi.org/10.1016/j.jclinepi.2013.06.018

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, *17*(1), 31–57.

Marcot, B. G., Steventon, J. D., Sutherland, G. D., & McCann, R. K. (2006). Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, *36*(12), 3063–3074. https://doi.org/10.1139/x06-135

Mulholland, J. A., Butler, A. J., Wilkinson, J. G., Russell, A. G., & Tolbert, P. E. (1998). Temporal and spatial distributions of ozone in Atlanta: Regulatory and epidemiologic implications. *Journal of the Air & Waste Management Association*, *48*(5), 418–426. https://doi.org/10.1080/10473289.1998.10463695

Office for National Statistics (2016a). Number of deaths from CVD, cancer, and diseases of liver by age group, year, month, and day of occurrence, English regions, deaths which occurred between 1981 and 2000. Published on 21st April 2016. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/adhocs/005639numberofdeathsfromcvdcanceranddiseasesofliverbyagegroupyearmonthanddayofoccurrenceenglishregionsdeathswhichoccurredbetween1981and2000

Office for National Statistics (2016b). Number of deaths from CVD, Cancer, and Diseases of Liver by age group, year, month, and day of occurrence, English Regions, deaths which occurred between 2001 and 2014. Published on the 2nd March 2016. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/adhocs/005433numberofdeathsfromcvdcanceranddiseasesofliverbyagegroupyearmonthanddayofoccurrenceenglishregionsdeathswhichoccurredbetween2001and2014

Office for National Statistics (2015). MYEDE Population Estimates for High Level Areas. Published on the 30th June 2015. A copy of the dataset (limited to residents over 60 years old) is available from this https://github.com/kehraProject/kehra/blob/master/data/PopulationEstimatesRegions1971_2014.rds

Pierce, D. (2015). ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files. R package version 1.15. Retrieved from https://CRAN.R-project.org/package=ncdf4

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.

Ryan, P. H., & LeMasters, G. K. (2007). A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation toxicology*, *19*(sup1), 127–133. https://doi.org/10.1080/08958370701495998

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, *35*(3), 1–22.

Scutari, M., Howell, P., Balding, D. J., & Mackay, I. (2014). Multiple quantitative trait analysis using Bayesian Networks. *GENETICS*, *198*(1), 129–137. https://doi.org/10.1534/genetics.114.165704

Turner, D. B. (1979). Atmospheric dispersion modelling: A critical review. *Journal of the Air Pollution Control Association*, *29*(5), 502–519. https://doi.org/10.1080/00022470.1979.10470821

Vitolo, C., Russell, A., & Tucker, A. (2017). Rdefra: Interact with the UK AIR Pollution Database from DEFRA. R package version 0.3.4. Retrieved from https://CRAN.R-project.org/package=rdefra, https://doi.org/10.5281/zenodo.838587

Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A., & Russell, A. (2017a). A multi-dimensional environment-health risk analysis system for the English regions. *19th EGU General Assembly, EGU2017, proceedings from the conference held 23-28 April, 2017 in Vienna, Austria* (11880 pp.). Retrieved from http://meetingorganizer.copernicus.org/EGU2017/EGU2017-11880.pdf

Vitolo, C., Scutari, M., Ghalaieny M., Tucker, A., & Russell, A. (2017b). MEHRA data and model for the English regions [Data set]. Zenodo. http://doi.org/10.5281/zenodo.838571

Vitolo, C., Russell, A., & Tucker, A. (2016). rdefra: Interact with the UK AIR Pollution Database from DEFRA. *The Journal of Open Source Software*, *1*(4). https://doi.org/10.21105/joss.00051