Proceedings of the 7th International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France - August, 2025
Paper No. 178
DOI: 10.11159/icsta25.178

Bayesian Network Modeling of Socio-Environmental Variables Supporting Health Policy Deliberation

Nicholas V. Scott[1], Sarah Jensen[2], and Tania Nur[2]
[1]Apogee Engineering, LLC, Science and Technology Group
2611 Commons Blvd., Beavercreek, OH, 45324, USA
nscott.gso@gmail.com; sarahjensen@franklincountyohio.gov; tanianur@franklincountyohio.gov
[2]Franklin County Public Health, Environmental Health
280 East Broad Street, Columbus, Ohio, 43215, USA

Abstract **-** Environmental health agencies face the critical challenge of understanding the interplay between air pollution and global socio-environmental quantifiers capturing land, water, and overall environmental health while operating under constrained budgets. Traditional analytical methods often fall short in capturing the complexity of these relationships. This study employs Bayesian belief network analysis to uncover crude structural dependencies and interactions between an array of environmental variables, offering a more holistic approach to data interpretation. Using data from the Center of Disease Control's tracking network (2005–2019), we examined 13 variables across 40+ Ohio zip codes, including land quality and air toxicity metrics (carbon tetrachloride, formaldehyde, benzene, acetaldehyde, and naphthalene). The Chow-Liu algorithm revealed land quality as a key parental node, with air toxicity variables acting as dependent children. A semi-inverse relationship emerged: higher land quality correlated with lower concentrations of formaldehyde and acetaldehyde, suggesting that land quality may serve as a coarse structural predictor of air pollution levels. To assess the possibility of basic predictive capability, we conducted hard instantiation simulations on land quality states. Drastic improvements in land quality reduced high-risk air toxicity levels by ~20%, while severe land quality degradation increased toxicity by ~30% for key pollutants. Notably, carbon tetrachloride levels remained unaffected, indicating potential data biases and highlighting other unique high mutual information-based pathways. These findings highlight the policy relevance of land quality as a leverage point for mitigating air pollution and how inference supported by Bayesian belief networks can support policy deliberation. Further research should address data gaps, particularly for pollutants like carbon tetrachloride, to refine inferences.

Keywords**:** Bayesian belief networks, correlation, air toxicity, land quality, Chow-Liu algorithm, pollutants, concentration, cancer

1. Introduction

Environmental health departments and organizations throughout the United States such as the Center for Disease Control (CDC) are often tasked with understanding the complex relationship between air toxicity and other socio environmental metrics under the constraint of tight budgets dictated by environmental public health policies. Focus is placed in this area because comprehension of such relationships impacts how health department distribute resources to combat the issue of clean air and water which affects many communities. The state of Ohio is one such place that ardently tries to deal with this issue through many avenues including the use of data analysis of measured and processed environmental variables [1]. Of particular importance now is not only the political issue of clean land, water, and air but also the more mundane aspect of how global parameterizations of these variables, which quantify overall environmental health, reflect more granular measurements of air toxicity affecting communities. A clear and valuable way of addressing this, paving the way towards improved environmental health, is through analysis of acquired data which seeks to go beyond simple classical methods of correlation between select environmental variables towards understanding systems of such variables as a gestalt.

Data supporting such nontraditional analyses are available from online databases created and maintained by the CDC [2]. Multivariate data sets allow for addressing two important questions. First is the comprehension of the structural relationship of measured and recorded environmental variables comprising the data array. In particular, with respect to achieving Bayesian structural clarity, there is a desire to determine if a global parent variable exists which is related to children variables establishing a binding cause-and-effect latent network structure. The second question is understanding if

expected state level changes within children environmental variables can be estimated from predetermined parent variable state level changes. Determining whether the crudest statistically simulated inferential changes are possible allows for gaining estimates of what environmental children variable changes should be expected if changes are prescribed in causal variables. Such understanding and insight allow for health administrators, who are concerned with influencing broad environmental policy, perspective into how granular measurements of air toxicity are potentially influenced by large scale environmental quantifiers which are the factors of great concern to them. Bayesian belief network (BBN) analysis allows for addressing these queries. While no inferences provided by a BBN is taken as sole definitive support for institutional policy change, having numerical analyses allows for more rigorous environmental deliberations than the use of suppositions backed by no quantifiable reasoning.

2. Data Structure and Methodology

A thirteen variable data set was used in this statistical analysis obtained from the CDC's data portal website which allows open-source data for public use [3]. The thirteen variables were selected: 1) overall environmental quality 2) environmental land quality, 3) environmental water quality, 4) air concentration for naphthalene 5) air concentration for formaldehyde 6) air concentration for carbon tetrachloride, 7) air concentration for benzene 8) air concentration for acetaldehyde, 9) mean cancer estimates due to airborne naphthalene, 10) mean cancer estimates due to airborne formaldehyde, 11) mean cancer estimates due to airborne carbon tetrachloride, 12) mean cancer estimates due to airborne benzene, and 13) mean cancer estimates due to airborne acetaldehyde. These variables consisted of numerical values for each variable where variables 1)-3) took on positive and negative values and variables 4) -13) took on only positive values. Data was curated and distilled into a m x n feature matrix with dimension of 88 x 13 where the n dimension designates the environmental variable. The m observations are the mean variable values over the time range spanning 2005-2019 for 88 zip codes spread over the state of Ohio and arranged in numerical order.

The raw data was compiled from the National Air Toxics Assessment (NATA) ran by the U.S. Environmental Protection Agency (EPA) which is a comprehensive evaluation of air toxins in the nation from 2005 to 2017. The Air Toxics Screening Assessment (AirToxScreen) began in 2017 provided data from 2017-2019. This compiled data was created for the purpose of gaining insight into which pollutants are potential targets for risk reduction activities. The data was also used to identify spatial locations of interest for further investigation, providing a starting point for local assessment and monitoring by programs and communities. With respect to cancer risk, the EPA calculated annual average cancer risk estimates to quantify the estimated lifetime probability of cancer from exposure to selected pollutants assessed in a geographic area. The EPA defined cancer risk as the probability of contracting cancer over the course of a lifetime assuming 70 years of continuous exposure. Annual average air concentration estimates were calculated by the EPA from outdoor air [3]. For the 2005-2017 NATA and the 2017-2019 AirToxScreen, cancer risks and air concentrations were calculated at the census tract level. The overall uncertainties and accuracy of the assessments varied from location to location and from pollutant to pollutant limiting the statistical inferences which can be made. All inferences and results therefore are not to be taken as absolutes but only provide and illustrate crude trends.

BBNs are probabilistic graphical models utilizing edges and nodes to model the joint probability distribution existing between a system of random variables [4]. They allow for statistical inferences to be made at random variable nodes when evidence is provided to one or more network nodes. Prior to statistical inference, network nodes along with nodal states need to be defined followed by structural learning which derives the directed acyclic graph (DAG) associated with the BBN. This step is concerned with exhuming the BBN topology from the data. Once the network is induced from data feature information, parameter learning can be performed which provides conditional probabilities relating different nodes [5]. The defined nodes and conditional probabilities in turn allow for statistical inference where the effects of evidence at one or more random variable nodes are propagated throughout the BBN to estimate its impact on other nodes.

BBN analysis was performed on the mean feature matrix using the software package Bayes Server manufactured by Bayes Server Ltd. which automates much of the statistical analysis including the Bayesian network structural learning and parameter learning. The Chow-Liu structural learning algorithm was implemented which is a global structural learning method that creates a single root initial network structure as the beginning of the network learning process [5]. The Chow-Liu algorithm assumes a tree structural model is appropriate to the data and seeks a skeletal structure consisting of a low number of dominant parental nodes which provides sub-dominant children nodes. A tree network grows iteratively in complexity using a structural model which contains a multivariate probability distribution expressed as a product of

conditional probability distributions associated with the nodes. The tree structure that best approximates the real distribution is found by minimizing the difference between the real data-based distribution and the tree approximation. This in turn is done by minimizing the mutual information between any two pairs of nodal variables [6]. The best network structure in other words is selected based on a score measuring how well the model represents the data. The direction of links is found using higher order dependency tests. Parameter learning provides numerical values to the nodal-edge structure allowing for statistical inference between nodes. The Chow-Liu algorithm also uses a relevance tree algorithm allowing for exact statistical inference rather than approximate inference [6].

3. Results

Fig. 1 shows the Sukiyama BBN layout for all 13 variables where no nodal evidential instantiations are implemented. Shown in green are nodes provided with parental node status before BBN structural learning was performed. Shown in blue are nodes provided with children node status before BBN structural leaning was performed. The BBN layout shows that all the nodes assigned children nodal status before structural learning appear as actual children nodal variables in the BBN. The BBN layout also shows that two of the three nodes assigned parental node status before structural learning appear as parental nodes. Structural learning therefore does exhume a parent-child latent structure in the multivariate data array. It is expected that the air toxin concentrations, air toxin cancer risks, and environmental health variables have normal and Poisson distributions due to the number of latent independent variables associated with these observed variables as well as the discrete low observation count for each variable. The histograms for each node are approximations of marginal distributions and have Gaussian and Poisson distributions consistent with the statistical process driving each variable.

The variable ENV-All is an overall environmental quality index encapsulating the factors of socio-demographics, and air, water, and land quality providing a crude summation of overall environmental health for Ohio counties. The variable ENV-All attempts to quantify areas with an increased burden of environmental impacts enabling counties to assess the drivers of poor environmental quality. It does not offer a measure of environmental quality at spatial and temporal scales that can inform individual-level adverse health outcomes. It is therefore a global variable of interest to many health administrators who work with health policy who have a need to understand large scale environmental health changes. The variables ENV-Land and ENV-water are also overall environmental quality indices for land and water respectively across Ohio counties.

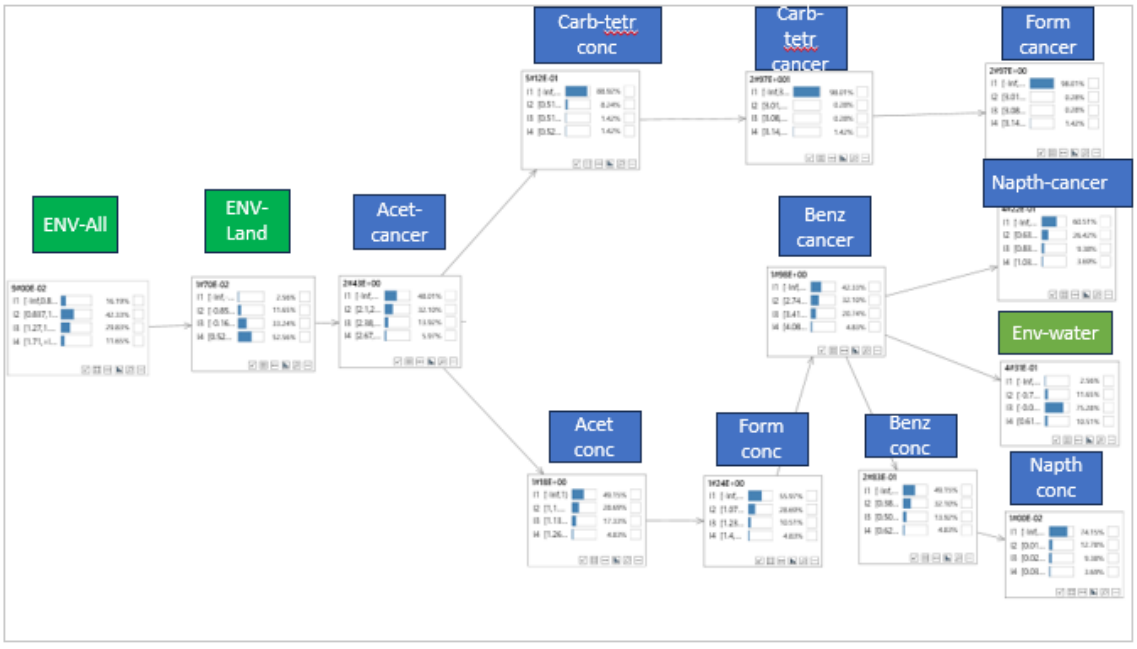

Figure 1: Sukiyama BBN network layout for 13 environmental variables estimated using the Chow-Liu algorithm. Each node contains a marginal distribution histogram with low state levels near the top and high state levels near the bottom. Arrow edges delineate covariant or mutual information connections. Green nodes are assumed parental node status labels before nodal structural learning was performed. Blue nodes are assumed children node status labels before nodal structural learning was performed.

Given the high amounts of uncertainty in the data variable array, the Chow-Liu algorithm was able to demonstrate that the environmental variables parameterizing overall environmental health, ENV-All, and environmental land quality, ENV-Land were parent variables giving 'birth' to air-toxin children variables. Mutual information is a way to quantify link strength in a BBN providing insight into which nodes share the strongest covariance and are the most amenable to covariant-based changes. The following nodes possessed strong mutual information link strengths: Env-All, Env-Land, Acet-cancer, Acet-conc, Form-conc, and Benz-cancer. The last four variables represent air toxicity parameterizing mean cancer estimates to airborne acetaldehyde, air concentration for acetaldehyde, air concentration for formaldehyde, and mean cancer estimates due to airborne benzene. The mutual information values for these nodes, which occupy the lower bifurcation pathway in the BBN, ranged from 0.15 – 0.9 suggesting that this pathway has the strongest cause-and-effect relationships. If crude statistical inferences regarding trends can be made using the relationships in the environmental variable array, the variables occupying the lower pathway in the BBN layout are the best candidates for performing this. For instance, based on all of the data spanning more than a decade from 2005-2019 and across 88 zip codes in Ohio, the BBN suggests that if any ties between socioeconomic variables and observed air toxicity variables exist, they most likely exist among these six variables. The upper bifurcation is characterized by very weak mutual information values suggesting that global inferences using ENV-All, ENV-Land, and air toxicity-based carbon tetrachloride cannot be made. More data is needed to understand this issue.

The BBN layout also shows the presence of air toxin couples for benzene and carbon tetrachloride parameterized using concentration and cancer risk which appear next to each other in the BBN layout, suggesting their high mutual information relationship. This is consistent with the idea that each member of an air toxin couple is quantifying the same air toxin variable albeit in different ways. The air toxin naphthalene parameterized using concentration and cancer risk appears furthest from the parent nodes of ENV-All and ENV-Land suggesting that this air toxin has the weakest mutual information relationship with the parent nodes. In addition, changes in the ENV-All and Env-Land variables produced negligible changes in air toxicity quantified using carbon-tetrachloride. The physical reasoning behind this is not clear at this point in time. The appearance of ENV-water as a child node all the way to the right in the BBN layout in addition to the very weak mutual information value between it and the node quantifying mean cancer estimates due to airborne benzene, which is on the order 0.09, demonstrates this variable has a very weak relationship to the other variables and is not a parental node as presupposed.

Instantiations in the BBN allow a crude sense of how changes produced in one variable via hard evidence affect other variables. Fig. 2 shows the same BBN in Fig.1 but where the ENV-Land variable has a hard evidential instantiation at the lowest interval which ranges over [-3 -0.859] and where the complete state intervals ranges over [-3 3]. (Here the units are arbitrary). This instantiation causes pronounced changes in the highest state intervals of the variables Acet-cancer, Acet-conc, Form-conc, and Benz-cancer all of which lie along the lower pathway after the bifurcation at the Acet-cancer node. The hard instantiation produces a 40% change in the Acet-cancer node, a 30% change in the Acet-conc node, a 30% change in Form-conc, and a 10% change in the Benz-cancer node. Fig. 3 shows the same BBN in Fig.1 but were the ENV-Land variable has a hard evidential instantiation at the highest interval which ranges over [0.522 3]. This instantiation causes pronounced changes in the lowest state intervals of the variables Acet-cancer, Acet-conc, Form-conc, and Benz-cancer. The hard instantiation produces a 25% change in the Acet-cancer and Acet-conc nodes, a 23% change in Form-conc, and a 3% change in the Benz-cancer node. The sociochemical reasons why these children nodes undergo these changes is unknown. The distillation of the data by the BBN however suggests how variables are possibly related and crudely show expected statistical changes in air toxicity based on projected changes in the socioeconomic environmental parental variable of Env-Land.

Environmental health departments possess a structure characterized by information flow from low level field technicians and scientists, responsible for the measurement and accrual of information closely affecting people, to higher level administrators responsible for initialization and implementation of broad health policy. Information also flows in the opposite direction where health policy-based logistics are delivered to field personnel responsible for interacting with the public in a variety of venues. This departmental structure is akin to the parent-children relationship of environmental variables demonstrated in the BBN. If a crude association is made where parental nodes are associated with high level administrators and children nodes with field technicians, the BBN may be viewed as a tool pertinent to environmental health departments' mission of improving environmental health policy. An example of how a BBN can be used in this way is furnished below.

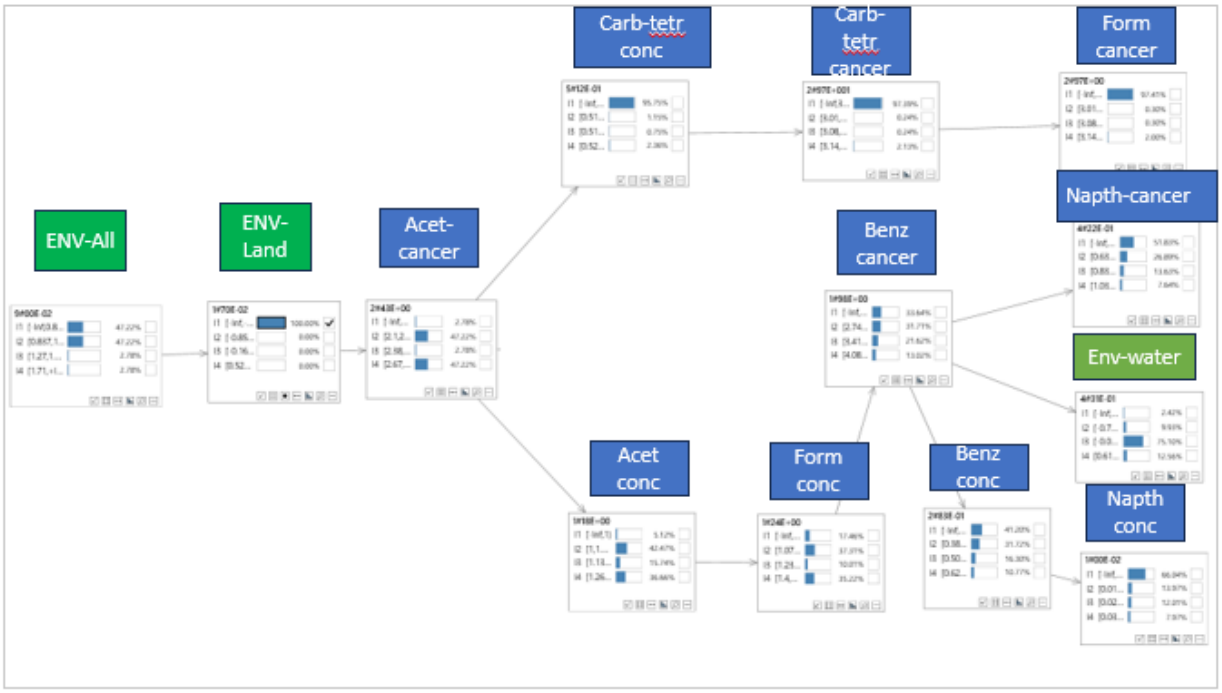

Figure 2: Sukiyama BBN layout of the BBN shown in Figure 1 where an instantiation is made at the lowest state level in ENV-All. The 4 state levels for ENV-All are [-3 -0.859], [-0.859 -0.169], [-0.169 0.522], and [0.522 3]. The 4 state levels for Acet-cancer and Acet-conc respectively are [0 2.1], [2.1 2.38, [2.38 2.67], and [2.67 4]; and [0 1.0], [1.0 1.13], [1.13 1.26], and [1.26 3]. The 4 state levels for Form-conc, and Benz-cancer respectively are [0 1.07], [1.07 1.23], [1.23 1.4], and [1.4 3]; and [0 2.74], [2.74 3.41], [3.41 4.08], and [4.08 6].

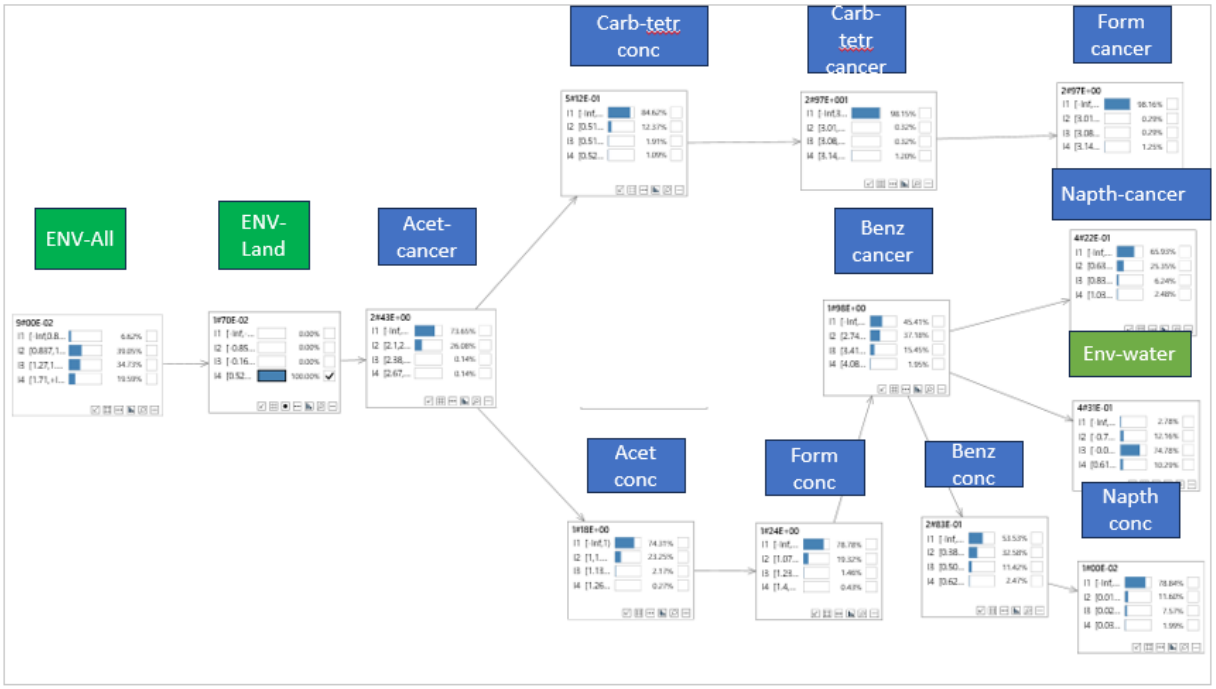

Figure 3: Sukiyama BBN layout of the BBN shown in Figure 1 where an instantiation is made at the highest state level in ENV-All. State levels for affected nodes explained in Figure 2.

BBN results show a coarse inverse relationship between changes in air toxicity nodes, and overall environmental quality and environmental land quality which is useful in shaping environmental policy. In addition, the BBN results show that when hard instantiations are provided at the lowest nodal state intervals of the parent node ENV-Land, substantial changes in the children variables sharing high mutual information ties are witnessed. These changes in children variables are much more substantial than when hard instantiations are instituted at the highest nodal state intervals of the ENV-Land variable. This is a potentially useful information facet distilled from the BBN for health administrators to be aware of since it suggests that environmental health policy measures directed towards environmental improvement in situations which are dire or bad are likely to show pronounced measurable changes in some air toxicity variables. In other words, when environmental health policy is directed towards improvement of situations which are least terrible, pronounced measurable changes in some air toxicity variables are less pronounced. The physical reasoning why these variables undergo the changes shown in the BBN simulations again is not completely understood. However, even with this uncertainty, the BBN crudely demonstrates expected change trends between variables sharing strong mutual information ties and therefore provides insight into how to properly shape policy.

4. Conclusion

In the state of Ohio, air quality has been a major concern as early as the 1970s and is an issue of renewed concern given recent disastrous events such as the train derailment in East Palestine, Ohio in 2023 [7, 8]. This accident, responsible for the release of large amounts of toxic chemicals into land, air, and water at such levels leading to evacuation orders, stresses the need for understanding the relationship between environmental quality variables and air toxicity concentration levels as well as air toxicity related to medical issues such as cancer. This age of artificial intelligence and machine learning allows for unprecedented ways and methods for the exploitation of readily available data, allowing organizations the ability to understand the relationship between environmental variables. BBNs are one such method for such comprehension. This work demonstrates how numerical data captured by health organizations can be distilled into useful information which can potentially guide environmental health deliberations. The intent of this work is not to suggest that BBNs alone can direct environmental health policy but to clearly demonstrate a rigorous way for presenting evidence that can affect how environmental health issues such are air toxicity are treated by health care practitioners.

References

[1] Franklin County Public Health Data Hub, https://fcph-data-hub-fca.hub.arcgis.com/pages/environmental-health, Accessed on 06/03/2025.

[2] WHO, Global Air Quality Guidelines: reinforcing the nexus between environment and health in the context of 'building forward better," *Science-policy dialogue, Air quality and Health*, Meeting Report, 27, March 2025, Bonn, Germany (online event), 14 October 2021, https://www.who.int/europe/publicatins/m/item/who-global-air-quality-guidelines-reinforcing-the-nexus-between-and-health, Accessed on 06/03/2025.

[3] Center for Disease Control National Environmental Public Health Tracking, https://ephtracking.cdc.gov/DataExplorer. Accessed on 06/03/2025.

[4] L. E. Sucar, *Probabilistic Graphical Models*. London, UK: Springer, 2015.

[5] A. Darwiche, *Modeling and Reasoning with Bayesian Networks.* New York, USA: Cambridge University Press, 2009.

[6] K. B Korb and A. E. Nicholson, *Bayesian Artificial Intelligenc*e. Florida, USA: CRC Press, 2010.

[7] Ohio Environmental Protection Agency website, https://dam.assets.ohio.gov/image/upload/epa.ohio.gov/Portals/47/facts/25%20years%20of%20protecting%20the%environment.pdf, Accessed on 06/03/2025.

[8] O. Oledeji, M. Saita, T. Mustapha, N. M/ Johnson, W. A. Chiu, I. Rusyn, A. L Robinson, and A. A. Presto, "Air Pollutant Patterns and Human Health Risk following the East Palestine, Ohio, Train Derailment", *Environmental Science and Technology Letters*, 10, pp. 680-685, doi: 10.1021/acs.estlett.3c00324, 2023.