Lichen-Based Bayesian Networks for Pollution Inference

Himanth Bobba Grant Cai Siena Okuno Tanishq Patil Michael Ye

Abstract

Air pollution monitoring within an area can be difficult and costly. This project leverages lichen as bioindicators to predict air pollution levels using probabilistic modeling. Two complementary Bayesian Network architectures using USFS lichen monitoring data are used.: the first incorporates a hidden node representing latent temporal environmental factors and the Expectation Maximization algorithm and the second uses Maximum Likelihood Estimation with fully observed environmental variables. By modeling the uncertainty in ecological responses and accommodating for real-world data limitations, these Bayesian Networks can serve as a powerful inference tool to model air pollution and aid in conservation efforts.

1 Problem Description

The era of industrialization at the end of the 19th century is defined as much by its technology as by industrial pollution. The passage of laws such as the Clean Air Act and Clean Water Act have significantly limited the release of pollutants into the atmosphere and water, but there are still concerns about the long-term effects of current pollution levels. As such, the need for easy and inexpensive monitoring methods has become apparent. However, direct analysis of pollutants within an area can be difficult and costly. Contaminants may be hard to measure depending on how they disperse within an environment, and bioavailability cannot be taken into account.

One method that has proven worthwhile is the use of certain organisms as bioindicators. The US Forest Service has been collecting data on lichen for this purpose since the 1970s. Lichen are particularly excellent specimens for bioindication. As epiphytes, they receive all of their nutrition and moisture from the air, which means they are especially sensitive to air pollution and changes in air quality. Their slow-growing but hearty nature also makes them highly likely to accumulate contaminants. By collecting data on tissue composition and abundance, scientists can monitor the health of local specimens and, by extension, forest health as a whole.

The goal of this project is to predict the probability of pollution within the environment based on the lichen species and tissue element analysis values of a sample. In this way, lichen can be used to evaluate ecological health without needing to directly measure pollutants. We have chosen to model the complex relationship between spatiotemporal data, lichen sample data, and pollution with a Bayesian Network in two ways; one containing a hidden node and one explicitly defining the spatiotemporal data relationships. Expectation Maximization and Maximum Likelihood Estimation are used to estimate the desired relationship within the networks respectively.

2 Data Sourcing and Processing

Our data comes from U.S. Forest Service (USFS) lichen biomonitoring programs, accessed via the NACSE "Lichen Air Quality" database exports (e.g., an air_lichen_query.csv file) and associated species lists. This data contains:

- Lichen community data at many field plots (which species were present, abundance, etc.).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- Lichen tissue chemistry for selected samples (element concentrations for metals and nutrients like copper, nitrogen, sulfur, etc.).
- Environmental variables for each sample or plot (region, elevation, slope, approximate collection date, and precipitation where available).
- Air pollution scores derived by USFS based on effects on lichen populations.

We used external reference sources (USFS documentation, lichen and pollution literature, bio-indicator literature) to inform our selection of a subset of this data with which to perform analysis.

At a broad level, we did three main kinds of preprocessing:

2.0.0.1 Removing missing data entries

Once we defined what data was to be used to build our model(s), sample entries that did not contain a complete instantiation for all nodes were removed. This reduced our training data from around 13,000 to a little under 8000, though it is important to note that some sample entries were extremely limited in what data was recorded.

2.0.0.2 Standardizing lichen species information

The scientific names from reference tables were parsed and standardized so that each species/taxon has a consistent identifier across all files. Then the standardized names were linked back to the element analysis data so that species-level patterns could be reliably analyzed. To ensure we had enough training data per species, species with under 500 samples were grouped into one species category.

2.0.0.3 Discretizing continuous variables into Bayesian Network–friendly categories

Since the network structure and CPTs are defined over a finite number of discrete states, all continuous variables were converted into categorical values. Tissue element concentrations were split into three categories (low, medium, high) based on percentiles, with $\frac{1}{3}$ of the data categorized into each label. The year a sample was collected was also split into three groups, while region relied on previously defined regions given in the original data.

While the original intention for discretizing air pollution score was to follow the USFS threshold values for 5 categories of air pollution, the dataset is not distributed equally between these categories and failed to learn the relationships between variables. Instead, air pollution scores were also binned by percentile into the categories of low, medium, and high. The lower bound of the "high" category roughly corresponds to USFS's threshold for if an area is affected by pollution (-0.11 and above). The split between "low" and "medium" categories does not directly correspond to a USFS pollution threshold, but can be thought of as "excellent" and "good", respectively.

Please see Appendix A for the a table describing each of the variables, their discretization, and the reasons for their inclusion in our models.

3 Modeling and Inference

The data can be arranged into a Bayesian Network to model dependencies between the environment, lichen tissue data, and the pollution found in the environment. The networks can be arranged into one utilizing complete data and one utilizing a hidden node to track unobserved temporal qualities about the environment.

For all equations described, the following notation holds true:

- T: total number of samples
- t: sample currently being evaluated
- $I(x, x_t)$: Indicator function. If $x_t = x$, then the result is 1. If not, the result is 0

3.1 Bayesian Network with Complete Data

This model (See Appendix B) is a simplified representation of the data without a hidden node, and is to be used when the variables in the Bayesian Network are fully observed. Without a hidden node, Field Collection Date and Region connect directly to Pollution in the Environment and Lichen Species. This model assumes that the direct temporal and regional information is sufficient to predict pollution levels. Due to its simpler structure and its use of complete, observed data, this model allows for direct interpretability.

3.1.1 Network Structure of Bayesian Network with Complete Data

Figure 2 depicts the graph structure of the Bayesian Network with Complete Data: Field Collection Date and Region serve as independent root nodes that connect directly to Pollution in the Environment and Lichen Species, without a hidden intermediary like with the model in Figure 1. Pollution in the Environment then also determines the lichen species and the tissue composition.

This Bayesian Network contains the following nodes:

- Observed Environment Variables:
  - R: Region
  - F: Field Collection Date
- Pe: Pollution in Environment
- Sp: Lichen Species present in the sample
- Tissue Composition Measurements (by dry weight of samples):
  - TNi: Tissue Composition - Nitrogen
  - TS: Tissue Composition - Sulfur
  - TP: Tissue Composition - Phosphorus
  - TPb: Tissue Composition - Lead
  - TCu: Tissue Composition - Copper
  - TCr: Tissue Composition - Chromium

3.1.2 Optimization with Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) estimates the CPTs that maximize the likelihood of observing the training data by computing empirical frequencies from complete observations. The advantages of utilizing a simplified MLE model here is that this calculation is computationally efficient and there are no convergence issues as there are with Expectation Maximization.

This means that for the root nodes we can calculate their likelihood as follows:

$$P(R = r) = \frac{\sum_{t=1}^{T} I(r, r_t)}{T}, \quad P(F = f) = \frac{\sum_{t=1}^{T} I(f, f_t)}{T},$$

The equations for Pollution $P(Pe \mid R, F)$, Lichen Species $P(Sp \mid F, Pe)$ and Tissue Composition $P(E \mid Sp, Pe)$ are as follows. E is used to represent all element nodes for tissue composition $\{TNi,\ TS,\ TP,\ TPb,\ TCu,\ TCr\}$.

$$P(Pe = p \mid R = r, F = f) = \frac{\sum_{t=1}^{T} I(p, p_t)\ I(r, r_t)\ I(f, f_t)}{\sum_{t=1}^{T} I(r, r_t)\ I(f, f_t)}$$

$$P(Sp = sp \mid R = r, F = f, Pe = pe) = \frac{\sum_{n:R^{(n)}=r, F^{(n)}=f, Pe^(n)=pe, Sp^{(n)}=sp,} 1}{\sum_{n:R^{(n)}=r, F^{(n)}=f, Pe^{(n)}=pe} 1}$$

$$P(T = t \mid Sp = sp, Pe = pe) = \frac{\sum_{n:Sp^{(n)}=sp, Pe^(n)=pe, T^{(n)}=t} 1}{\sum_{n:R^{(n)}=r, F^{(n)}=f} 1}$$

We are looking to infer the pollution level given the lichen species and tissue composition. Thus we have

$$= P(Pe|T, Sp) = \frac{P(Pe, T, Sp)}{P(T, Sp)}$$

$$= \frac{\sum_{r',f'} P(Pe, T, r', f', Sp)}{\sum_{r',f',pe'} P(T, r', f', Sp, pe')}$$

$$\text{numerator} = \sum_{f',r'} P(r')P(f')P(Pe|r', f')P(Sp|r', f')P(T|Sp, Pe)$$

$$= P(T|Sp, Pe) \sum_{f'} P(f') \sum_{r'} P(r')P(Pe|r', f')P(Sp|r', f')$$

$$f_1 = \sum_{r'} P(r')P(Pe|r', f')P(Sp|r', f')$$

$$= P(T|Sp, Pe) \sum_{f'} P(f')f_1$$

$$f_2 = \sum_{f'} P(f')f_1$$

$$= P(T|Sp, Pe)f_2$$

$$\text{finally, we have } P(Pe|T, Sp) = \alpha P(T|Sp, Pe)f_2$$

3.2 Bayesian Network with Hidden Node

The rationale for introducing a hidden node to model hidden spatiotemporal aspects of the environment is twofold: firstly, despite the wealth of columns in the data about the environment in which the lichen were found in (region, elevation, etc.), there still could be some aspects of the environment not directly captured or observed in the dataset, such as microclimate. Addressing these in a hidden node will allow for the model to be more robust in its modeling of the relationship between lichen, their environment, and pollutants by acting as a latent variable that captures these other hidden aspects. Consequently, this model assumes that the effect of the temporal aspects of the environment on other nodes in the graph can be approximated by this hidden node. Secondly, some data in the dataset itself is lacking. The EM algorithm in conjunction with the hidden node structure handles incomplete data by inferring missing values in the E-step. This contrasts with case deletion or imputation that could introduce bias.

3.2.1 Network Structure of Bayesian Network with Hidden Node

For this Bayesian Network (See Appendix C), all nodes are the same as for the previous model besides the addition of a hidden node H. This node separates our region/environment variables from the lichen species and air pollution score. The goal of this node is to learn latent spatial temporal aspects such as seasonal patterns or regional mineral baselines. The presence of this node means that we cannot naively use a Maximum Likelihood Estimate like we would for a network with complete data without some additional work. Therefore, we will opt to learn our CPTs through the Expectation Maximization (EM) algorithm. For each sample we aim to first calculate the probability of our hidden variable (the relevance of our regional data) given our other observed features. This probability is our posterior. Then, in the Maximization step, we want to update our CPTs using the posterior we calculated in the E step. Essentially, we will use our estimated expectation to fill in the value of our hidden node. The ultimate goal of using EM for this task and introducing our hidden variable is to give our model some flexibility in terms of how it weights the relevance of our region based features.

3.2.2 Optimization with Expectation Maximization

3.2.2.1 E-step: Obtain posterior for hidden node

In the Expectation Step, we compute the posterior distribution of the hidden variable H given the visible nodes $F,\ R,\ Pe\ Sp$. For a value of $H = h$, we define:

$$P(H = h \mid F = f, R = r, Pe = p, Sp = s) = \frac{P(h, f, r, p, s)}{\sum_{h' \in H} P(f, r, h', p, s)} = \gamma(h, f, r, p, s)$$

Since all tissue composition nodes are D-separated from the hidden node by the visible nodes Pe and Sp, they do not need to be included in the posterior. The joint distribution of these terms can be expanded using the CPTs. For brevity's sake, this equation is referred to as $\gamma(h, f, r, p, s)$.

3.2.2.2 M-step: Update CPTs

In the Maximization Step, we update all CPTs using the expected counts from the E-step.
The root nodes R and F do not rely on H, so they can be updated using the standard MLE frequency counts. This equation describes that, where X represents each respective node:

$$P(X = x) = \frac{\sum_{t=1}^{T} I(x, x_t)}{T}$$

The hidden node h is updated by aggregating the expected counts of $H = h$ for parent-node values $F = f, R = r$, then normalizing. It is updated with the following equation:

$$P(H = h \mid F = f, R = r) \leftarrow \frac{\sum_{t=1}^{T} I(f, f_t)\ I(r, r_t)\ \gamma(h, f, r, p_t, s_t)}{\sum_{t=1}^{T} I(f, f_t)\ I(r, r_t)}$$

The pollution node Pe is updated by summing the expected counts across all samples with the following equation:

$$P(Pe = p \mid H = h) \leftarrow \frac{\sum_{t=1}^{T} I(p, p_t)\ \gamma(h, f_t, r_t, p_t, s_t)}{\sum_{t=1}^{T} \gamma(h, f_t, r_t, p_t, s_t)}$$

Similar to pollution, the lichen species node Sp is updated using expected counts of parent $H = h$ along with parent $Pe = p$. It uses the following equation:

$$P(Sp = p \mid H = h, Pe = p) \leftarrow \frac{\sum_{t=1}^{T} I(s, s_t)\ I(p, p_t)\ \gamma(h, f_t, r_t, p, s_t)}{\sum_{t=1}^{T} I(p, p_t)\ \gamma(h, f_t, r_t, p, s_t)}$$

The tissue composition nodes $\{TNi,\ TS,\ TP,\ TPb,\ TCu,\ TCr\}$ are separated D from the hidden node, and therefore can follow the standard MLE formula. It is calculated with the following equation, where X represents each respective node:

$$P(X = x \mid Sp = s, Pe = p) = \frac{\sum_{t=1}^{T} I(x, x_t)\ I(s, s_t)\ I(p, p_t)}{\sum_{t=1}^{T} I(s, s_t)\ I(p, p_t)}$$

3.2.2.3 Running EM

The E and M steps are repeated in turn for every value of every node until the log-likelihood of the data converges or a maximum number of iterations is reached.

4 Results and Discussion

4.1 Compare and Contrast of Different Network Structures

We explored two different belief networks for this project, one with and one without a hidden node representing unobserved temporal qualities of the environment. These networks required us to use the Maximum Likelihood Estimate and Expectation Maximization algorithms.

4.2 Quantitative Results

To evaluate both the MLE and EM implementations, we utilized the accuracy of predictions on the test set (20% of the dataset.) We calculated our predictions by taking our discretized air pollution score buckets and selecting the one that has the highest probability given the rest of our data using the CPTs of our model. Then, we can calculate the accuracy of our model on the test set by taking all of our predictions for each data point and determining the ratio of number of correct predictions to total number of samples. (For a table comparing the models, see Appendix D.)

The MLE model predicting $P(Pe \mid Sp, T)$ achieved an accuracy of 52.2%. Overall, our EM model appears to perform significantly better than our implementation of Maximum Likelihood Estimation. We can see that we are able to achieve peak accuracy scores of around **72.4%**, and it seems that trying different numbers of iterations: {25, 50, 100, 200, 300} did not improve our prediction accuracy any more than what we achieved at 200 iterations.

During training time, we also tracked the log-likelihood of our data in order to verify that we are learning. We can observe that our log-likelihood appears to plateau as we increase the number of iterations, indicating that we are close to convergence. We also attempted to vary the number of buckets for hidden nodes that we had in our belief network. Setting the number of hidden nodes from 2 to any of {3, 5} did not result in an increase in accuracy. Increasing the number of hidden nodes to 3 yielded an accuracy score of 57.5% and increasing to 5 nodes yielded a higher accuracy score of 70.8% which was still less than what we achieved with 2 hidden nodes.

4.3 Qualitative Insights

Discuss what aspects of the data does the model capture or miss, and do the learned CPTs or latent states align with intuition?

4.4 Convergence and Scalability

For MLE, this calculation is performed on the observed data; thus, it scales linearly with the size of the dataset. More data will obviously yield a more accurate reflection of the data. As a result, our choice of configurations are limited since adjusting configurations for the MLE model would require shifting away from the network structure we believe is ideal for capturing the node relationships of our dataset. Similarly, beyond changing the network structure for the EM model, we are limited to only adjusting the initialization values of our model.

For evaluating our EM implementation, we relied on accuracy score of predictions on our test set(which is 20% of our dataset). For reference, we calculate our predictions by taking our discretized air pollution score buckets and selecting the one that has the highest probability given the rest of our data using the CPTs of our model. Then, we can calculate the accuracy of our model on the test set by taking all of our predictions for each data point and determining the ratio of number of correct predictions to total number of samples. During training time, we also track the log-likelihood of our data in order to verify that we are learning (Appendix E).
Overall, our EM model appears to perform significantly better than our implementation of Maximum Likelihood Estimation. We can see that we are able to achieve peak accuracy scores of around **72.4%**, and it seems that trying different numbers of iterations: {25, 50, 100, 200, 300} did not improve our prediction accuracy any more than what we achieved at 200 iterations. We can also see that our log-likelihood appears to plateau as we increase the number of iterations, indicating that we are close to convergence. We also attempted to vary the number of buckets for hidden nodes that we had in our belief network. Setting the number of hidden nodes from 2 to any of {3, 5} did not result in an increase in accuracy. Increasing the number of hidden nodes to 3 yielded an accuracy score of 57.5% and increasing to 5 nodes yielded a higher accuracy score of 70.8% which was still less than what we

achieved with 2 hidden nodes.

5 Conclusion

We faced several limitations when developing our models. While we strove to simulate a scientifically accurate belief network using nodes derived from the data columns of our source, there were instances where we were unsure how to incorporate certain nodes like abundance. Furthermore, we were unable to use every column from the dataset due to sparsity and bad distribution. We had originally intended to employ far more parent nodes that capture environmental aspects like precipitation, elevation, and more, but we were limited by how sparse they were in the dataset and ultimately chose to not include them. Potential extensions for our work would include building a more extensive belief network that brings together additional nodes that we excluded from this experiment.

6 Reflections & Contributions

Member contributions were as follows:
Tanishq: I worked on drafting the equations for the inference and updates for our CPTs using EM, implementing EM on our cleaned dataset, implementing log likelihood as a method of evaluation, and testing/evaluating the EM model. Through the process, I gained a much stronger understanding of how and where EM can be applied and gained experience in the process of translating belief network structure and CPT equations into code.
Michael: I worked on the Maximum Likelihood Estimate algorithm, specifically the inference for the bayesian network and the evaluation functions. On the report, I wrote the parts for the MLE inference and the results pertaining to MLE. I learned a lot from the project, from the subject material regarding lichen and the different environmental impacts of pollutants. Additionally, I got the opportunity to help create a belief network (and implement it in code) that models a real-world model with tangible applications as measuring air pollution is a relevant problem for today.
Grant: I worked on the data discretization, formulating the model section of the paper, and MLE implementation, primarily formulating the equations and implementing the counts and CPTs. I learned overall how to apply the concept of MLE to real-world datasets and implement it into code.
Himanth: I worked on the data analyzation and processing, and I was also in charge of creating the Latex files for the first two milestones. I learned about how hard it is to work with a real world database in comparison to ones made specifically for homework.
Siena: I worked on drafting the EM equations, modeling both BNs, filtering and binning the dataset, and implementing the naive predictor. I learned a lot in terms of how to appropriately model a real-world system and how to work with messy data.

References

[1] https://www.nzdr.ru/data/media/biblio/kolxoz/P/PGp/Hill%20M.K.%20Understanding%20Environmental%20Pollution%20(draft,%203ed.,%20CUP,%202010)(ISBN%200521518660)(O)(602s)_PGp_.pdf

[2] https://www.envchemgroup.com/understanding-environmental-pollution-element-by-element.html

[3] https://www.researchgate.net/figure/Periodic-table-of-environmental-impacts-colored-according-fig3_263708668

[4] https://gis.nacse.org/lichenair/index.php?page=cleansite

[5] https://www.sciencedirect.com/science/article/abs/pii/S0045653520316301

[6] https://internationalcopper.org/sustainable-copper/about-copper/copper-in-the-environment/

[7] https://www.sciencedirect.com/science/article/abs/pii/S030147972100236X

[8] https://www.sciencedirect.com/science/article/pii/S0045653524009214

[9] https://gis.nacse.org/lichenair/index.php?page=airpollution#metals

[10] https://www.atsdr.cdc.gov/toxprofiles/tp26-c1.pdf

[11] https://www.nature.com/scitable/knowledge/library/bioindicators-using-organisms-to-measure-environmental-impacts-16821310/

[12] https://www.sciencedirect.com/science/article/pii/S2950395723000012

[13] https://www.fs.usda.gov/rm/pubs_rm/rm_gtr224.pdf

[14] https://oceanservice.noaa.gov/education/tutorial_pollution/02history.html

[15] https://plants.usda.gov/downloads

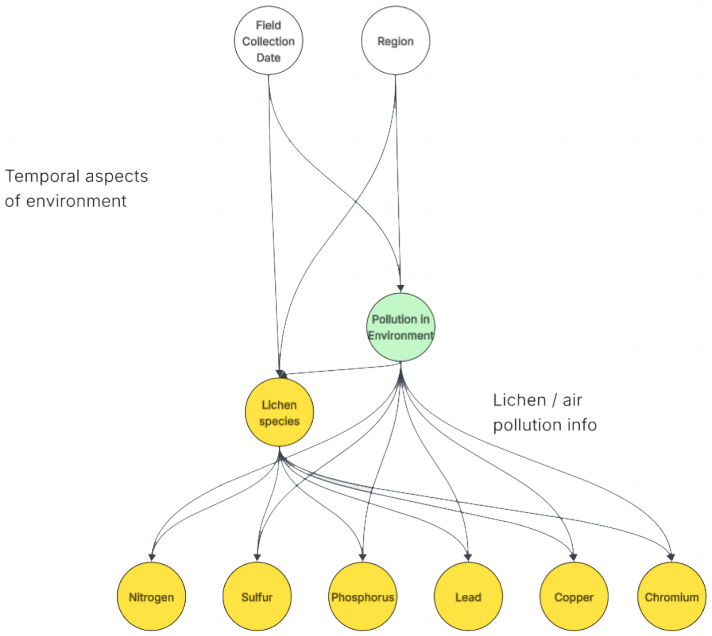
7 Appendix

A Variable Discretization and Categorization

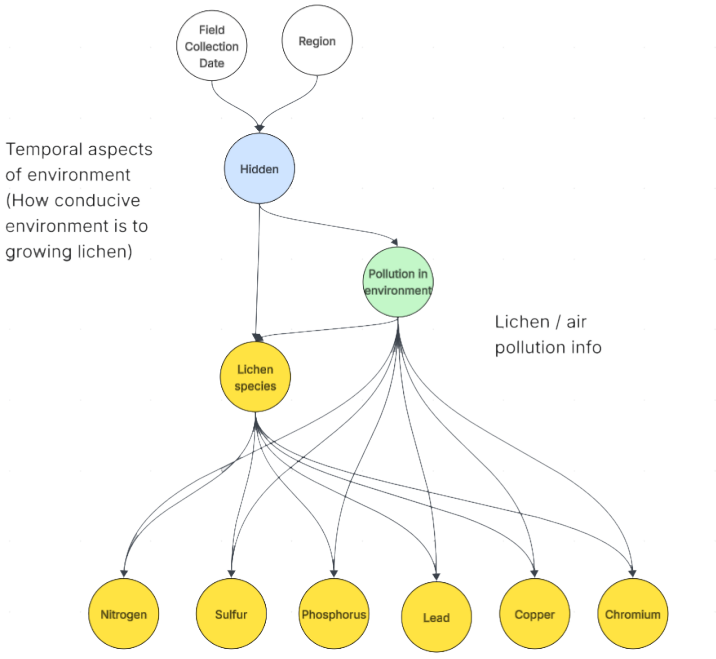
Table 1: Variable discretization and categorization

Name	Label	Justification for inclusion
Region	1-4, 5, 6, 7, 8-9, 10	Locations for a sample are split into 10 regions across the US. The amount of a given element naturally available varies across regions.
Field Collection Date	Before 1995 1995-2005 2005 to present	Pollution levels can change over time within an area from events like the passage of environmental protection laws or the opening of a mining operation nearby.
Air pollution score	Low ($<^{-} 0.46$) Medium($^{-}0.46 -^{-} 0.14$) High (> -0.139)	This is a categorized score of ambient air quality, often determined by proximity to major industrial or traffic sources. This is our query node.
Lichen Species	Alectoria sarmentosa Flavoparmelia caperata Hypogymnia inactiva Letharia vulpina Platismatia glauca Other	Lichen species can react very differently to different amounts of specific pollutants in the environment. By including this category, some of these relationships may be accounted for.
Tissue Composition: Nitrogen	Low (< 0.472) Medium ($0.472 - 0.72$) High (> 0.72)	Nitrogen is a naturally found element that is essential to life. However, human usage of synthetic fertilizers and fossil fuels generate harmful nitrogen oxides (smog/greenhouse gasses).
Tissue Composition: Sulfur	Low (< 0.048) Medium ($0.048 - 0.071$) High (> 0.071)	While beneficial in limited qualities, large amounts introduced to the environment through burning coal and heavy oil can lead to acid rain.
Tissue Composition: Phosphorus	Low (< 639.16) Medium ($639.16 - 928$) High (> 928)	Phosphorus is found in commercial fertilizers. It is known to cause effects such as limited plant growth in freshwater lakes and deformities in frogs.
Tissue Composition: Lead	Low (< 2.86) Medium ($2.86 - 3.739$) High (> 3.739)	Lead (and lead compounds) are among the greatest carcinogenic hazards to human health. Lead is exposed in mining operations or from leaded paint.
Tissue Composition: Copper	Low (< 2.17) Medium($2.17 - 4.535$) High (> 4.535)	Copper is a naturally occurring heavy metal that is essential to life but toxic in high quantities. Only a small amount of copper in an environment is typically bio-available, which means directly observing the environmental levels is not an accurate measure of its effects. Copper can accumulate due to mining, fertilizer, manufacturing, etc.
Tissue Composition: Chromium	Low ($< .6$) Medium ($0.6 - 1.19$) High (> 1.19)	This trace element is one of the most toxic heavy metals. While the naturally occurring form is essential to life, even small amounts of Cr produced through industrial processes has extremely detrimental effects.

B Bayesian Network Structure with Complete Data



C Bayesian Network with Hidden Node



D Comparison of Pollution Element (Pe) Prediction Accuracy Across Different Bayesian Network Modeling Techniques

Model Description	Estimation Method	Accuracy (%)
Full Network Inference $P(Pe \mid Sp, T)$	MLE (Counting)	52.2
Full Network Inference (with Latent Variables/Missing Data)	EM Algorithm	72.4

E Log Likelihood