
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.).

- 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 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 $P(Pe | R, F)$, Lichen Species $P(Sp | F, Pe)$ and Tissue Composition $P(E | Sp, Pe)$ are as follows. E is used to represent each element node for tissue composition $\{TNi, TS, TP, TPb, TCu, TCr\}$.

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

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

$$P(E = e | Sp = s, Pe = p) = \frac{\sum_{t=1}^T I(e, e_t) I(s, s_t) I(p, p_t)}{\sum_{t=1}^T I(s, s_t) I(p, p_t)}$$

We are looking to infer the pollution level given the lichen species and tissue composition. We represent this as

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

Using variable elimination to break this down into factors, we have

$$\begin{aligned} P(Pe|T, Sp) &= \frac{P(Pe, T, Sp)}{P(T, Sp)} = \frac{\sum_{r', f'} P(Pe, T, r', f', Sp)}{\sum_{r', f', pe'} P(T, r', f', Sp, pe')} \\ &= \alpha \sum_{f', r'} P(r') P(f') P(Pe|r', f') P(Sp|r', f') P(T|Sp, Pe) \\ &= \alpha P(T|Sp, Pe) f_2 \end{aligned}$$

where

$$\begin{aligned} f_1 &= \sum_{r'} P(r') P(Pe|r', f') P(Sp|r', f') \\ f_2 &= \sum_{f'} P(f') f_1 \end{aligned}$$

α is a normalizing constant such that the resulting probabilities sum to 1 i.e. $\sum_q P(q|E) = 1$

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:

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

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. As a reminder, we compute our log likelihood as:

$$\log L = \sum_{i=1}^N \log \left[\sum_{h=1}^H P(h \mid r_i, f_i) P(Pe_i \mid h) P(sp_i \mid h, Pe_i) \prod_{t \in \mathcal{T}} P(t_{ik} \mid Pe_i, sp_i) \right]$$

Once we have learned our CPTs, we want to determine a relationship between the element analysis and tissue data from our lichen. We do this using our CPTs to calculate:

$$P(Pe = pe \mid T=t, Sp = sp) = \frac{\sum_h P(H = h) P(Pe = pe \mid H = h) \prod_{t \in \mathcal{T}} P(t = t \mid Pe = pe, sp)}{\sum_{pe'} \sum_h P(H = h) P(Pe = pe' \mid H = h) \prod_{t \in \mathcal{T}} P(t = t \mid Pe = pe', sp)}$$

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.) Our baseline model that predicts the median air pollution score for all samples achieved an accuracy of 33.6%. 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 only when we include Region and Field Collection date. We are able to achieve peak accuracy scores of around **72.4%** if we compute $P(Pe \mid R, F, Sp, \{t \in T\})$. On the other hand, computing only $P(Pe \mid Sp, \{t \in T\})$ (as described above) yields only an accuracy score of **50.2%**. Overall, on the primary objective of predicting $P(Pe \mid T, Sp)$, our EM model appears perform marginally worse than Maximum Likelihood Estimation(MLE), something that can likely be attributed to the smoothing factors we introduce to prevent deterministic probability values (equal to 0).

During training time, we also tracked the log-likelihood of our data in order to verify that we are learning (see Appendix E). We can observe that our log-likelihood appears to plateau as we increase the number of iterations, indicating that we are close to convergence. Further increasing or decreasing the number of training iterations do not affect our final prediction accuracy. We also attempted to vary the number of hidden nodes that we had in our belief network. Setting the number of hidden nodes to all of $\{2, 3, 5\}$ did not result in an increase or decrease in accuracy. Rather, we saw that the log-likelihood converged/collapsed slightly faster with more hidden nodes.

4.3 Qualitative Insights

Taking a look at the CPTs learned from our MLE model (Appendix F), we can quickly observe that region 10 historically has much better air quality than region 6. This aligns with our intuition as region 10 represents Alaska, which is know to one of the lowest air pollution areas in the US due to its low population density whereas region 6 represents the Pacific Northwest which has a higher population density and more industrial activity. Our model appears to have missed the temporal aspect of the data as the CPTs indicate that air pollution has gotten worse in Alaska and the PNW over time, which is counter to the general trend of air pollution in the US decreasing over time due to environmental regulations.

The CPTs learned from our EM model (Appendix G) also indicate clearly that Region is the dominant driver behind our hidden node H. In particular, Region 10 has a strong correlation (99%+) to one of our hidden node buckets, and Region 6 has an equally strong correlation to another. Furthermore, we can see that our hidden node (and therefore region) heavily influence the prediction for air pollution, where $H=0$ indicates pollution is low 80% of the time, and $H=1$ has a more even spread across pollution values. This is consistent with what we saw in the CPTs learned via MLE. We also see that, as expected, specific region(s) are tied heavily to our species nodes, like in region 10 ($H=0$) which is correlated to high probability density contributions from Species 1 (*Alectoria Sarmentosa*) across pollution levels.

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 and help the model converge on the true value. 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.

As described above, our implementation of EM was able to consistently converge in roughly 200 iterations. This can be visibly seen in the plot of log likelihood in the Appendix. We also attempted to vary the number of hidden nodes in our belief network to test if increasing the number of learnable parameters(scaling up) would give us a better accuracy. However, training with $\{2, 3, 5\}$ hidden

nodes did not result in an increase in accuracy. With regards to computational cost, EM performs linearly with respect to the size of our dataset, meaning that this approach would likely be feasible to scale up. Increasing the amount of data should allow for our learned CPTs to be more precise and generalized. Furthermore, introducing more sample diversity to our data should potentially make scaling up our belief network a more viable strategy. This is because a larger dataset would potentially contain more complex trends requiring more hidden variables to model effectively.

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 poor data 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 data exploration and 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 design and implement a belief network 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\(draft,%203ed.,%20CUP,%202010\)\(ISBN%200521518660\)\(0\)\(602s\)_PGp_.pdf](https://www.nzdr.ru/data/media/biblio/kolxoz/P/PGp/Hill%20M.K.%20Understanding%20Environmental%20Pollution%20(draft,%203ed.,%20CUP,%202010)(ISBN%200521518660)(0)(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-to-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>
- [16] <https://www.epa.gov/air-trends/air-quality-national-summary>
- [17] <https://www.usnews.com/news/best-states/rankings/natural-environment/air-water-quality/urban-air-quality>

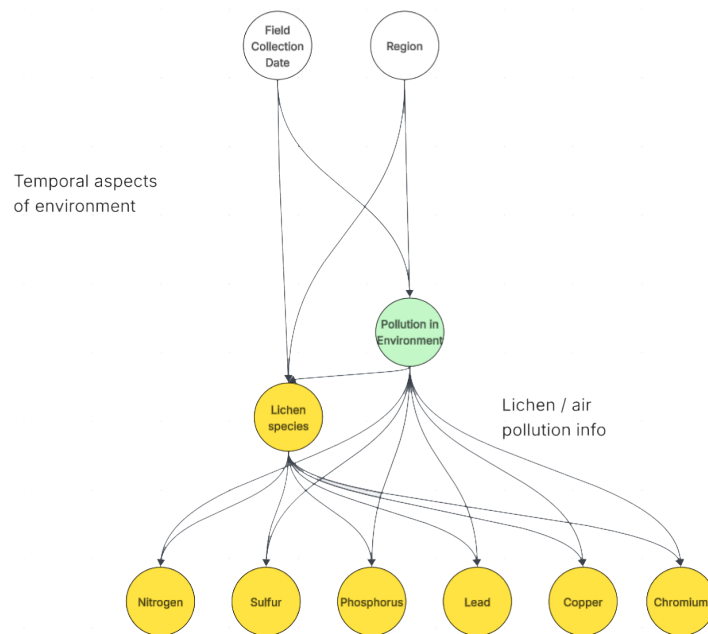
7 Appendix

A Variable Discretization and Categorization

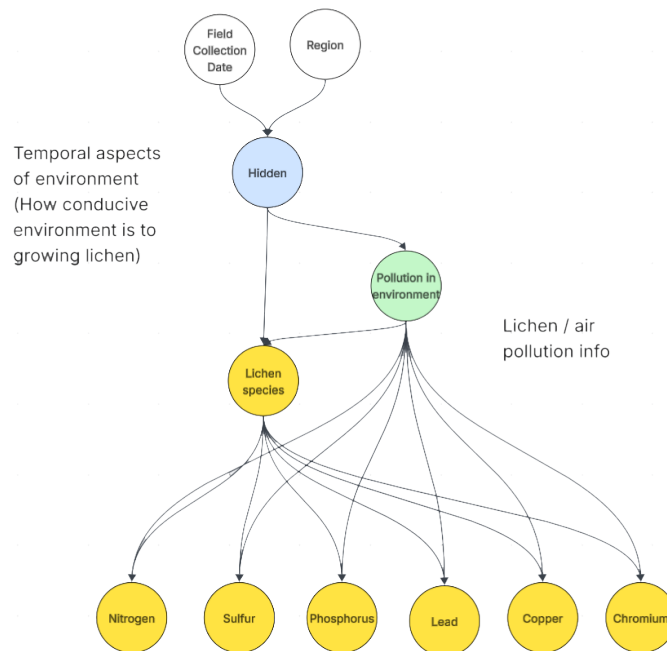
Table 1: Variable discretization and categorization

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(1) Flavoparmelia caperata(2) Hypogymnia inactiva(3) Letharia vulpina(4) Platismatia glauca(5) Other(6)	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 quantities, 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



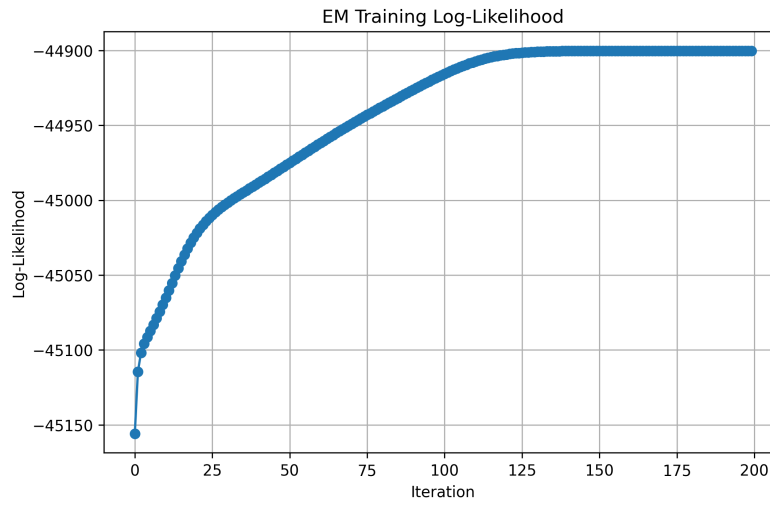
C Bayesian Network with Hidden Node



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

Model Description	Estimation Method	Accuracy (%)
Baseline Predictor	Predict Median	33.6
Full Network Inference $P(Pe \mid Sp, T)$	MLE (Counting)	52.2
Full Network Inference (with Latent Variables)	EM Algorithm	50.2

E EM Log Likelihood



F MLE CPTs

Table 2: Marginal Probability Table $P(\text{Region})$

Region	Probability
10	0.0446
6	0.9554
Total	1.0000

Table 3: Marginal Probability Table $P(\text{Year})$

Year	Probability
before 1995	0.2201
1995–2005	0.6141
2005–present	0.1658
Total	1.0000

Table 4: CPT: $P(\text{Pollution} \mid \text{Region, Field Collection Date})$

Region	Time Period	Low	Medium	High
6	Before 1995	0.3398	0.3541	0.3061
6	1995–2005	0.2942	0.3651	0.3407
6	2005–Present	0.3174	0.2895	0.3931
10	Before 1995	0.8333	0.1282	0.0385
10	1995–2005	0.8387	0.0565	0.1048
10	2005–Present	0.7222	0.2407	0.0370

Table 5: CPT: $P(\text{Species} \mid \text{Region 10, Field Collection Date, Pollution})$

Time Period	Pollution	Species	Probability
before 1995	High	Species 1	1.0000
before 1995	Low	Species 1	1.0000
before 1995	Medium	Species 1	1.0000
1995-2005	High	Other	0.1538
1995-2005	High	Species 1	0.3077
1995-2005	High	Species 5	0.5385
1995-2005	Low	Other	0.2115
1995-2005	Low	Species 1	0.5962
1995-2005	Low	Species 5	0.1923
1995-2005	Medium	Other	0.2857
1995-2005	Medium	Species 1	0.2857
1995-2005	Medium	Species 5	0.4286
2005-present	High	Species 1	0.5000
2005-present	High	Species 5	0.5000
2005-present	Low	Other	0.2821
2005-present	Low	Species 1	0.3846
2005-present	Low	Species 5	0.3333
2005-present	Medium	Other	0.4615
2005-present	Medium	Species 1	0.0769
2005-present	Medium	Species 5	0.4615

Table 6: CPT: $P(\text{Species} \mid \text{Region 6, Field Collection Date, Pollution})$

Time Period	Pollution	Species	Probability
before 1995	High	Other	0.5069
before 1995	High	Species 1	0.1074
before 1995	High	Species 3	0.0882
before 1995	High	Species 4	0.1074
before 1995	High	Species 5	0.1901
before 1995	Low	Other	0.3871
before 1995	Low	Species 1	0.2953
before 1995	Low	Species 3	0.0174
before 1995	Low	Species 4	0.1117
before 1995	Low	Species 5	0.1886
before 1995	Medium	Other	0.3310
before 1995	Medium	Species 1	0.3143
before 1995	Medium	Species 3	0.0548
before 1995	Medium	Species 4	0.0500
before 1995	Medium	Species 5	0.2500
1995-2005	High	Other	0.3624
1995-2005	High	Species 1	0.1044
1995-2005	High	Species 3	0.1337
1995-2005	High	Species 4	0.1242
1995-2005	High	Species 5	0.2752
1995-2005	Low	Other	0.2927
1995-2005	Low	Species 1	0.2537
1995-2005	Low	Species 3	0.0260
1995-2005	Low	Species 4	0.1848
1995-2005	Low	Species 5	0.2428
1995-2005	Medium	Other	0.2472
1995-2005	Medium	Species 1	0.2625
1995-2005	Medium	Species 3	0.0942
1995-2005	Medium	Species 4	0.0821
1995-2005	Medium	Species 5	0.3140
2005-present	High	Other	0.3088
2005-present	High	Species 1	0.0793
2005-present	High	Species 3	0.1020
2005-present	High	Species 4	0.2550
2005-present	High	Species 5	0.2550
2005-present	Low	Other	0.2421
2005-present	Low	Species 1	0.1965
2005-present	Low	Species 3	0.0421
2005-present	Low	Species 4	0.2386
2005-present	Low	Species 5	0.2807
2005-present	Medium	Other	0.2038
2005-present	Medium	Species 1	0.2462
2005-present	Medium	Species 3	0.0577
2005-present	Medium	Species 4	0.2462
2005-present	Medium	Species 5	0.2462

Table 7: CPT: $P(\text{Nitrogen} \mid \text{Pollution, Species})$

Species	Pollution	Low	Medium	High
Other	low	0.1688	0.2704	0.5608
Other	medium	0.2189	0.3116	0.4694
Other	high	0.0867	0.1706	0.7427
Species 1	low	0.8021	0.1821	0.0158
Species 1	medium	0.6972	0.2766	0.0262
Species 1	high	0.5255	0.3980	0.0765
Species 3	low	0.0667	0.7333	0.2000
Species 3	medium	0.1355	0.6774	0.1871
Species 3	high	0.1166	0.3049	0.5785
Species 4	low	0.1040	0.4899	0.4060
Species 4	medium	0.1123	0.4973	0.3904
Species 4	high	0.0073	0.2711	0.7216
Species 5	low	0.5208	0.4560	0.0231
Species 5	medium	0.5299	0.4384	0.0317
Species 5	high	0.1975	0.4053	0.3971

Table 8: CPT: $P(\text{Sulfur} \mid \text{Pollution, Species})$

Species	Pollution	Low	Medium	High
Other	low	0.1652	0.3702	0.4646
Other	medium	0.2150	0.3925	0.3925
Other	high	0.0825	0.2797	0.6378
Species 1	low	0.7968	0.1226	0.0806
Species 1	medium	0.7383	0.1290	0.1327
Species 1	high	0.6888	0.2143	0.0969
Species 3	low	0.1333	0.5333	0.3333
Species 3	medium	0.1032	0.5484	0.3484
Species 3	high	0.0538	0.2735	0.6726
Species 4	low	0.2383	0.6141	0.1477
Species 4	medium	0.2513	0.5508	0.1979
Species 4	high	0.1209	0.4872	0.3919
Species 5	low	0.4676	0.4213	0.1111
Species 5	medium	0.3979	0.4577	0.1444
Species 5	high	0.1193	0.3457	0.5350

Table 9: CPT: $P(\text{Phosphorus} \mid \text{Pollution, Species})$

Species	Pollution	Low	Medium	High
Other	low	0.2196	0.3122	0.4682
Other	medium	0.1519	0.2189	0.6292
Other	high	0.0867	0.2531	0.6601
Species 1	low	0.8021	0.1751	0.0228
Species 1	medium	0.7140	0.2299	0.0561
Species 1	high	0.5306	0.3316	0.1378
Species 3	low	0.0222	0.2222	0.7556
Species 3	medium	0.0452	0.4516	0.5032
Species 3	high	0.0359	0.3049	0.6592
Species 4	low	0.6409	0.2987	0.0604
Species 4	medium	0.4492	0.3369	0.2139
Species 4	high	0.3553	0.3700	0.2747
Species 5	low	0.2384	0.6111	0.1505
Species 5	medium	0.2711	0.5616	0.1673
Species 5	high	0.1687	0.3868	0.4444

Table 10: CPT: $P(\text{Lead} \mid \text{Pollution, Species})$

Species	Pollution	Low	Medium	High
Other	low	0.5372	0.3376	0.1252
Other	medium	0.4142	0.2505	0.3353
Other	high	0.2685	0.2573	0.4741
Species 1	low	0.5937	0.3082	0.0981
Species 1	medium	0.4561	0.4037	0.1402
Species 1	high	0.5204	0.4031	0.0765
Species 3	low	0.1333	0.2000	0.6667
Species 3	medium	0.0581	0.1677	0.7742
Species 3	high	0.0179	0.2287	0.7534
Species 4	low	0.5034	0.3489	0.1477
Species 4	medium	0.4278	0.4064	0.1658
Species 4	high	0.3809	0.4029	0.2161
Species 5	low	0.1782	0.3542	0.4676
Species 5	medium	0.1004	0.3856	0.5141
Species 5	high	0.1173	0.3519	0.5309

Table 11: CPT: $P(\text{Copper} \mid \text{Pollution, Species})$

Species	Pollution	Low	Medium	High
Other	low	0.2577	0.4338	0.3085
Other	medium	0.1834	0.3629	0.4536
Other	high	0.0825	0.3580	0.5594
Species 1	low	0.7723	0.1208	0.1068
Species 1	medium	0.7289	0.1664	0.1047
Species 1	high	0.6888	0.1582	0.1531
Species 3	low	0.0000	0.5111	0.4889
Species 3	medium	0.0000	0.4258	0.5742
Species 3	high	0.0135	0.2556	0.7309
Species 4	low	0.6745	0.1611	0.1644
Species 4	medium	0.6952	0.1497	0.1551
Species 4	high	0.5604	0.3150	0.1245
Species 5	low	0.1736	0.5347	0.2917
Species 5	medium	0.1056	0.5387	0.3556
Species 5	high	0.0679	0.4671	0.4650

Table 12: CPT: $P(\text{Chromium} \mid \text{Pollution, Species})$

Species	Pollution	Low	Medium	High
Other	low	0.3938	0.3448	0.0000
Other	medium	0.2229	0.4280	0.3491
Other	high	0.1091	0.3203	0.5706
Species 1	low	0.8827	0.0841	0.0333
Species 1	medium	0.8654	0.1047	0.0299
Species 1	high	0.7806	0.1633	0.0561
Species 3	low	0.1333	0.5111	0.3556
Species 3	medium	0.0452	0.5097	0.4452
Species 3	high	0.0314	0.2780	0.6906
Species 4	low	0.4329	0.4933	0.0738
Species 4	medium	0.4813	0.4492	0.0695
Species 4	high	0.3040	0.4542	0.2418
Species 5	low	0.1134	0.4514	0.4352
Species 5	medium	0.0370	0.4419	0.5211
Species 5	high	0.0453	0.2860	0.6687

G EM CPTs

Table 13: Marginal Probability Table $P(\text{Region})$

Region	Probability
10	0.0446
6	0.9554
Total	1.0000

Table 14: Marginal Probability Table $P(\text{Field Collection Date})$

Year	Probability
before 1995	0.2201
1995–2005	0.6141
2005–present	0.1658
Total	1.0000

Table 15: Conditional Probability Table $P(\text{Chromium (binned)} \mid \text{Pollution, Species})$

Pollution	Species	low	medium	high
low	Other	0.3938	0.3448	0.2613
low	Species 1	0.8826	0.0841	0.0333
low	Species 3	0.1335	0.5110	0.3555
low	Species 4	0.4329	0.4933	0.0739
low	Species 5	0.1134	0.4514	0.4352
medium	Other	0.2229	0.4280	0.3491
medium	Species 1	0.8654	0.1047	0.0299
medium	Species 3	0.0452	0.5096	0.4451
medium	Species 4	0.4813	0.4492	0.0696
medium	Species 5	0.0370	0.4419	0.5211
high	Other	0.1091	0.3203	0.5706
high	Species 1	0.7805	0.1633	0.0562
high	Species 3	0.0314	0.2780	0.6905
high	Species 4	0.3040	0.4542	0.2418
high	Species 5	0.0453	0.2860	0.6687

Table 16: Conditional Probability Table $P(\text{Copper (binned)} \mid \text{Pollution, Species})$

Pollution	Species	low	medium	high
low	Other	0.2577	0.4338	0.3085
low	Species 1	0.7723	0.1209	0.1068
low	Species 3	0.0002	0.5110	0.4888
low	Species 4	0.6745	0.1611	0.1644
low	Species 5	0.1736	0.5347	0.2917
medium	Other	0.1834	0.3629	0.4536
medium	Species 1	0.7289	0.1664	0.1047
medium	Species 3	0.0001	0.4258	0.5741
medium	Species 4	0.6951	0.1498	0.1551
medium	Species 5	0.1056	0.5387	0.3556
high	Other	0.0825	0.3580	0.5594
high	Species 1	0.6887	0.1582	0.1531
high	Species 3	0.0135	0.2556	0.7309
high	Species 4	0.5604	0.3150	0.1246
high	Species 5	0.0679	0.4671	0.4650

Table 17: Conditional Probability Table $P(\text{Lead (binned)} \mid \text{Pollution, Species})$

Pollution	Species	low	medium	high
low	Other	0.5372	0.3376	0.1252
low	Species 1	0.5937	0.3082	0.0981
low	Species 3	0.1335	0.2001	0.6664
low	Species 4	0.5033	0.3490	0.1477
low	Species 5	0.1783	0.3542	0.4676
medium	Other	0.4142	0.2505	0.3353
medium	Species 1	0.4561	0.4037	0.1402
medium	Species 3	0.0581	0.1678	0.7741
medium	Species 4	0.4278	0.4064	0.1658
medium	Species 5	0.1004	0.3856	0.5141
high	Other	0.2685	0.2573	0.4741
high	Species 1	0.5204	0.4031	0.0766
high	Species 3	0.0180	0.2287	0.7533
high	Species 4	0.3809	0.4029	0.2161
high	Species 5	0.1173	0.3519	0.5309

Table 18: Conditional Probability Table $P(\text{Phosphorous (binned)} \mid \text{Pollution, Species})$

Pollution	Species	low	medium	high
low	Other	0.2196	0.3122	0.4682
low	Species 1	0.8021	0.1751	0.0228
low	Species 3	0.0224	0.2223	0.7553
low	Species 4	0.6409	0.2987	0.0604
low	Species 5	0.2384	0.6111	0.1505
medium	Other	0.1519	0.2189	0.6292
medium	Species 1	0.7140	0.2299	0.0561
medium	Species 3	0.0452	0.4516	0.5032
medium	Species 4	0.4492	0.3369	0.2139
medium	Species 5	0.2711	0.5616	0.1673
high	Other	0.0867	0.2532	0.6601
high	Species 1	0.5306	0.3316	0.1378
high	Species 3	0.0359	0.3049	0.6591
high	Species 4	0.3553	0.3700	0.2747
high	Species 5	0.1687	0.3868	0.4444

Table 19: Conditional Probability Table $P(\text{Sulfur (binned)} \mid \text{Pollution, Species})$

Pollution	Species	low	medium	high
low	Other	0.1652	0.3702	0.4646
low	Species 1	0.7968	0.1226	0.0806
low	Species 3	0.1335	0.5332	0.3333
low	Species 4	0.2383	0.6141	0.1477
low	Species 5	0.4676	0.4213	0.1111
medium	Other	0.2150	0.3925	0.3925
medium	Species 1	0.7383	0.1290	0.1327
medium	Species 3	0.1033	0.5483	0.3484
medium	Species 4	0.2514	0.5508	0.1979
medium	Species 5	0.3979	0.4577	0.1444
high	Other	0.0825	0.2797	0.6377
high	Species 1	0.6887	0.2143	0.0970
high	Species 3	0.0538	0.2736	0.6726
high	Species 4	0.1209	0.4872	0.3919
high	Species 5	0.1194	0.3457	0.5350

Table 20: Conditional Probability Table $P(\text{Nitrogen (binned)} \mid \text{Pollution, Species})$

Pollution	Species	low	medium	high
low	Other	0.1688	0.2704	0.5608
low	Species 1	0.8021	0.1821	0.0158
low	Species 3	0.0668	0.7331	0.2001
low	Species 4	0.1040	0.4899	0.4060
low	Species 5	0.5208	0.4560	0.0232
medium	Other	0.2189	0.3116	0.4694
medium	Species 1	0.6972	0.2766	0.0262
medium	Species 3	0.1355	0.6774	0.1871
medium	Species 4	0.1123	0.4973	0.3904
medium	Species 5	0.5299	0.4384	0.0317
high	Other	0.0867	0.1706	0.7426
high	Species 1	0.5255	0.3979	0.0766
high	Species 3	0.1166	0.3049	0.5784
high	Species 4	0.0074	0.2711	0.7216
high	Species 5	0.1975	0.4053	0.3971

Table 21: Conditional Probability Table $P(\text{Species} \mid H, \text{Pollution})$

H	Pollution	Other	Species 1	Species 3	Species 4	Species 5
0	low	0.1881	0.6646	0.0000	0.0000	0.1472
0	medium	0.2861	0.4463	0.0003	0.0002	0.2671
0	high	0.1300	0.4608	0.0005	0.0005	0.4082
1	low	0.3749	0.0008	0.0433	0.2867	0.2942
1	medium	0.2578	0.2612	0.0853	0.1030	0.2928
1	high	0.3875	0.0894	0.1225	0.1499	0.2507

Table 22: Conditional Probability Table $P(\text{Pollution} \mid H)$

H	Pollution	Probability
0	low	0.8047
0	medium	0.1276
0	high	0.0677
1	low	0.2222
1	medium	0.3884
1	high	0.3894

Table 23: Conditional Probability Table $P(H \mid \text{Region}, \text{Year})$

Region	Year	$P(H = 0)$	$P(H = 1)$
10	before 1995	0.9998	0.0002
10	1995–2005	0.9995	0.0005
10	2005–present	0.9809	0.0191
6	before 1995	0.2061	0.7939
6	1995–2005	0.1377	0.8623
6	2005–present	0.1095	0.8905