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 or categories derived by USFS from lichen community composition.

We also use external reference sources (USFS documentation, lichen and pollution literature, and plant/lichen species lists) to interpret and standardize species names and to choose meaningful thresholds for pollution and environmental buckets.

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

Cleaning and standardizing lichen species information Parse and normalize scientific names from reference tables so that each species/taxon has a consistent identifier across all files. Join those standardized names back to the lichen plot and tissue chemistry data so that species-level patterns can be analyzed reliably.

Discretizing continuous variables into Bayesian Network–friendly categories Air pollution score is converted into a categorical pollution level (e.g., low/medium/high) using USFS-based thresholds. Continuous environmental variables like collection date are binned into a small number of buckets (e.g., "before 1995") using domain-informed breakpoints. Tissue element concentrations (e.g., copper or nitrogen in lichen tissue) are also bucketed (e.g., "background / elevated / high") to make conditional probability tables tractable.

These steps are crucial because: If species names are inconsistent or misaligned across tables, we'd mix different taxa and distort the relationship between lichen communities, tissue chemistry, and pollution. Standardizing and joining species information ensures that when the model learns "this species tends to occur at high nitrogen sites" or "this tissue concentration pattern is associated with metal pollution," it's actually talking about the same organism. The network structure and CPTs are defined over finite states, so we must discretize continuous features (pollution scores, elevation, tissue concentrations, etc.) into meaningful categories. Doing this with USFS thresholds and literature-based breakpoints keeps the categories ecologically interpretable (e.g., what counts as "high" copper or "high" pollution from a lichen-bioindicator perspective).

I'll show the exact dataset information on Piazza, when I have time to write it up.

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.

3.1 Bayesian Network with Complete Data

This model 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.

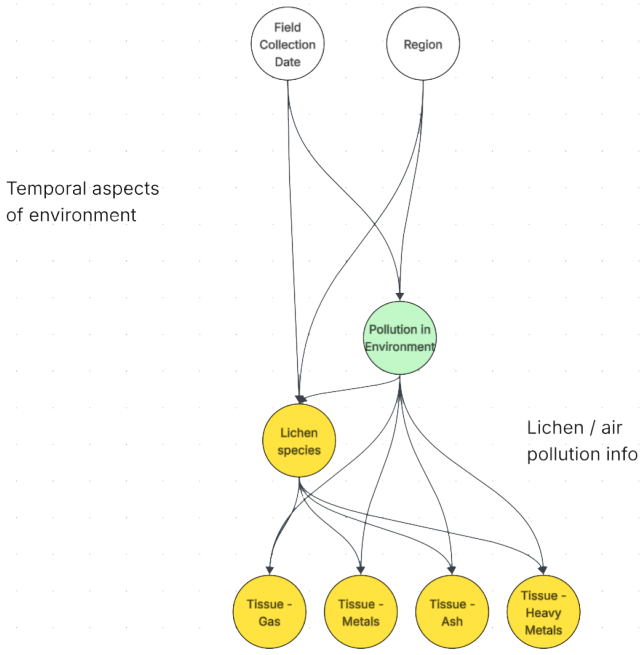

Figure 1: Bayesian Network Structure with Complete Data

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:

$$P(R = r) = \frac{\sum_{n:R^{(n)}=r} 1}{N}, \quad P(F = f) = \frac{\sum_{n:F^{(n)}=f} 1}{N},$$

and the equations for pollution $P(Pe \mid R, F)$, Lichen Species $P(Sp \mid F, Pe)$ and Tissue Composition $P(T \mid Sp, Pe)$ are the following:

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

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

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

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

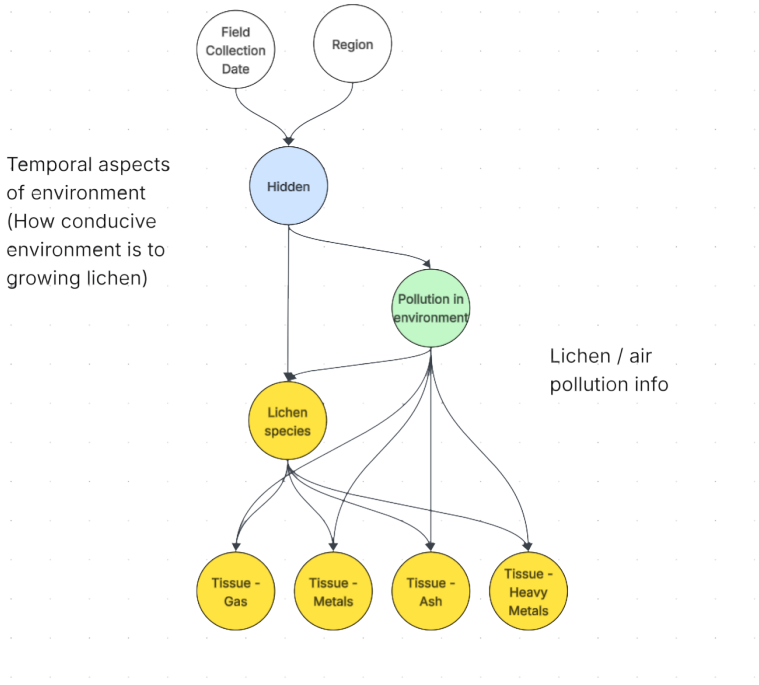

Figure 2: Bayesian Network Structure with Hidden Node

Our Bayesian Network contains the following nodes:

- Observed Environment Variables:
  - R: Region
  - F: Field Collection Date
- H: Hidden Environmental Factor (latent temporal aspects like seasonal pattern, microclimate, etc.)
- Pe: Pollution in Environment (latent given H)
- Sp: Lichen Species present in the sample
- Tissue Composition Measurements:
  - Tg: Tissue-Gas (gas content in lichen tissue)
  - Tm: Tissue-Metals (metal concentrations in lichen tissue)
  - Ta: Tissue-Ash (ash content in lichen tissue)
  - Th: Tissue-Heavy Metals (heavy metal concentration in lichen tissue)

For this Bayesian Network, we have a hidden node that separates our region/environment variables from the lichen species and air pollution score. The goal of the node is to learn a variable that represents spatiotemporal aspects of the environment to predict air pollution level. 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

In the Expectation Step, we compute the probability distribution over the hidden variable H given the pollution level Pe for each data point and the current parameter estimates and observed data. For each sample n, we define:

$$\gamma^{(n)}(h, pe) = P(H = h \mid Pe = pe, \text{ observed data}^{(n)})$$

This is calculated using Bayes' rule; the full equation multiplies $P(h \mid R^{(n)}, F^{(n)}$ by the likelihood $P(pe \mid h)$, and by the likelihoods of all observed measurements $P(Sp^{(n)} \mid h, pe)$, $P(Tg^{(n)} \mid Sp^{(n)}, pe)$, $P(Tm^{(n)} \mid Sp^{(n)}, pe)$, $P(Ta^{(n)} \mid Sp^{(n)}, pe)$ and $P(Th^{(n)} \mid Sp^{(n)}, pe)$. The denominator normalizes over all possible combinations of h and pe.

In the Maximization Step, we update all CPTs using the expected counts from the E-step. The expected count of hidden state h and pollution level pe is defined as:

$$N(h, pe) = \sum_n \gamma^{(n)}(h, pe)$$

which represents how many times the hidden state was h and the pollution was pe across all samples.

Different nodes are updated differently based on whether they involve H or not.

Beginning with the Environmental Nodes R and F, since they do not depend on H or Pe, they can be updated using standard MLE frequency counts.

$$P(R = r) = \frac{\sum_{n:R^{(n)}=r} 1}{N}, \quad P(F = f) = \frac{\sum_{n:F^{(n)}=f} 1}{N},$$

The hidden state's CPT is updated by aggregating the expected counts γ over samples with matching parent values of R and F, the normalizing. Let $\pi_H = (R, F)$:

$$P(H = h \mid \pi_H) = \frac{\sum_{n:\, \pi_H^{(n)}=\pi_H} \sum_{pe} \gamma^{(n)}(h, pe)}{\sum_{n:\, \pi_H^{(n)}=\pi_H} \sum_{h'} \sum_{pe} \gamma^{(n)}(h', pe)}$$

Pollution $P(Pe \mid H)$ is updated by summing expected counts across all samples for each (h, pe) pair:

$$P(Pe = pe \mid H = h) = \frac{\sum_n \gamma^{(n)}(h, pe)}{\sum_n \sum_{pe'} \gamma^{(n)}(h, pe')}$$

For each lichen species value, we weight by the posterior probability of each hidden state configuration:

$$P(Sp = sp \mid H = h, Pe = pe) = \frac{\sum_{n:Sp^{(n)}=sp} \gamma^{(n)}(h, pe)}{\sum_n \gamma^{(n)}(h, pe)}$$

Finally, for the tissue nodes, each tissue variable $T \in \{Tg, Tm, Ta, Th\}$ depends on the observed species and pollution level. We marginalize over h by summing the posteriors:

$$P(T = t \mid Sp = sp, Pe = pe) = \frac{\sum_{n:T^{(n)}=t} \sum_h \gamma^{(n)}(h, pe)}{\sum_n \sum_h \gamma^{(n)}(h, pe)}$$

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

4 Results and Discussion

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. 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.

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

Blank

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