Graphical Models of Air Pollution in Beijing

David Lu
UCI
lud19@uci.edu

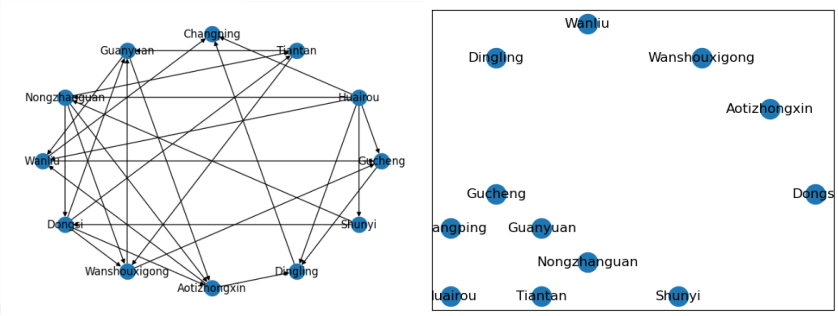
Abstract

Air pollution in Beijing, China, is a critical problem requiring innovative analysis. The focus of our project is to apply graphical models - specifically, Bayesian Networks (BN) and Markov Random Fields (MRF) - to a deep-dive study of regional pollution data. These graphical models are pivotal, as they enable the representation of complex dependencies and allow us to model relationships between the pollution levels in different areas. Beginning with BN, we will analyze the pollution correlations, and subsequently, using MRFs, we will explore the interactions between regions.

1 Introduction

Many people in Beijing, China are bothered by air pollution for years. (Murtaugh, 2022) It is still short of effective measures to control air pollution in Beijing, China. Different regions in Beijing have different air pollution levels, but the levels in these regions seem to correlate. If we can figure out the relationship of air pollution across different regions in Beijing, it may shed light on the air pollution issue in Beijing and inspire environmental researchers to come up with more effective remedies for air pollution.

To explore the dependencies of the pollution levels in different regions, we plan to construct a Bayesian Network (BN) and a Markov Random Field (MRF) on the time-series air pollution data in these regions. The example Bayesian Network (BN) and Markov Random Field (MRF) crafted in this project are shown in Figure. 1. Each node represents the daily average PM2.5 by district and is discretized.



(a) Learned Structure of a BN (b) MRF with No Edges

Figure 1: Example BN and MRF

Several studies have explored the application of BN in weather prediction. Cofino and his team Cofino et al. (2002) has investigated the causal relationships of rainfall in 100 cities of Spain with the aid of the BN. In their research, they constructed the graphical structure utilizing the K2 algorithm

Preprint. Under review.

and leveraged the data to derive the directed weighted edges. Aye Nander Nandar (2009) applied BN to explore the spatial dependencies among the meteorological variables for rainfall and temperature prediction over Myanmar. One group in Japan (Dindar et al., 2018) developed a weather forecasting system with an artificial neural network and BN.

The contribution of this project is two-folded. First, while numerous studies applied graphical models in precipitation or other meteorological variables, there is sparse research on air pollution in Beijing, which is still a headache in Beijing. Building graphical models can uncover the correlation of air pollution in different regions of Beijing, which inspires more creative and efficient plans for air pollution in Beijing. Moreover, by incorporating other meteorological factors, such as wind direction and air pressure, into graphical models, we gain a deeper understanding of underlying causes for air pollution and comprehend the dynamics of pollution dispersion.

2 Approach

2.1 Bayesian Network

2.1.1 K2 Metric and Hill-Climbing Search

In our study, we plan to employ structure learning on a BN. Each node in the network will represent the daily average PM2.5 level for a different region. Our chosen method of structure learning will be the Hill-Climbing algorithm, combined with the K2 score as our scoring method. The K2 score was selected for its advantageous bias towards simpler models, and the fewer assumptions it makes about the data, thus avoiding overfitting. This approach allows us to balance model complexity and learn the structure from the data, which is essential in situations with limited data. On the other hand, the Hill-Climbing algorithm is an efficient heuristic approach for navigating the large search space of potential network structures, providing an effective compromise between computational efficiency and quality of the resulting model. The algorithm essentially looks for the local maximum K2 metric.

We will focus on the K2 scoring (metric). Borgelt and Kruse explain that the K2 metric is a specific Bayesian Dirichlet metric. The scoring function can be written as:

$$K2(A|\gamma(A)) = \prod_{j=1}^{q} \frac{(r-1)!}{(N_{.j}+r-1)!} \prod_{i=1}^{r} N_{ij}! \tag{1}$$

where A serves as the child node, and $\gamma(A)$ represents the set of parent attributes associated with A. The attribute A can take on r different values, while q represents the number of unique combinations or instantiations of its parent attributes. Nij refers to the count of instances where attribute A takes on its i-th value, with the parents instantiated using the j-th value combination, and $N_{.j}$ is computed by summing up the counts Nij for all i (Borgelt and Kruse, 2001). And by using the Hill-Climbing algorithm, we try to maximize the K2 score from an initially fully-disconnected BN to find the best structure according to the data.

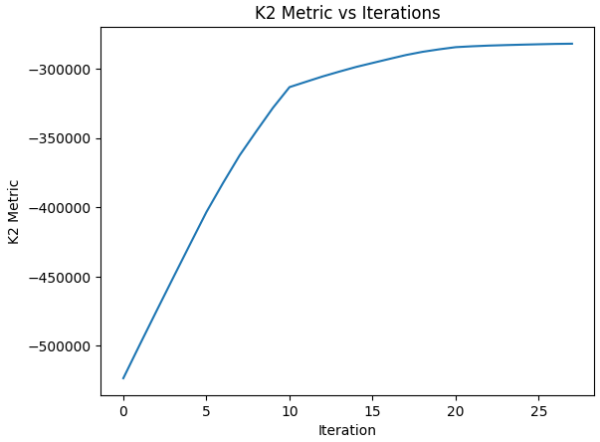
2.1.2 Implementation detail

We constructed the Bayesian Network (BN) utilizing the BayesianNetwork module from the pgmpy library. To facilitate structure learning, we made slight adjustments to the Hill-Climbing algorithm, enabling it to monitor the K2 metric throughout each iteration. Our initial setup for the starting graph was a fully disconnected BN. As illustrated in Figure 2, there's a notable progressive increase in the K2 score with each iteration, eventually leading to its convergence. The BN implementation can be viewed by going to the url:`https://github.com/david5010/Graphical-Model-for-Pollution-Data.git`.

2.2 Markov Random Field

Other than Bayesian Network (BN), we explore the pairwise correlations across different regions in Beijing with pairwise Markov Random Field (MRF).

Figure 2: Example of the learning objective improving



Pairwise MRF In this project, we focus on learning pairwise MRF from the training set of air pollution data in Beijing. The pairwise MRFs crafted in this project all contain 12 nodes, representing 12 regions in Beijing. Each node represents air pollution category in one region at each day. The value in each node n, denoted as x_n, is either 0 or 1, with 0 representing healthy and 1 unhealthy. The daily air pollution in 12 regions can be recorded as $\{x_1, x_2, x_3, ..., x_{12}\}$. The training set has 1168 days of air pollution condition in Beijing. Then we have 1168 samples in training set, each sample is air pollution condition of 12 regions in one day. To parameterize such a MRF model, we will use θ_s for each node s $\in \nu$, and θ_{st} for each edge $(s,t) \in \varepsilon$. The joint distribution of **daily** air pollution in 12 regions can be expressed as (Of note, the parameterization is borrowed from the HW2):

$$p(x|\theta) = exp\{\sum_{s\in\nu} \theta_s x_s + \sum_{st\in\varepsilon} \theta_{st} x_s x_t - \Phi(\theta)\}$$
$$\Phi(\theta) = \sum_{x} (\theta_s x_s + \theta_{st} x_s x_t)$$

2.2.1 Log-likelihood of MRF model without L1 regularization

From above equations, we can compute the log-likelihood of a single observation given $\theta_s \in \nu$ and $\theta_{st} \in \varepsilon$. Therefore, the log-likelihood of MRF model given the training set (training set has 1168 days) can be defined as:

$$logp(X|\theta) = \sum_{l=1}^{1168} (\sum_{s\in\nu} \theta_s x_s^l + \sum_{st\in\varepsilon} \theta_{st} x_s^l x_t^l - \Phi(\theta)) \tag{2}$$

2.2.2 Training objective

L1 regularization Without L1 regularization, by optimizing the log-likelihood in equation(2) a generated MRF model is very likely to be fully connected. With L1 regularization, some pairwise edges may be removed from the MRF model, because L1 regularization will impose penalties on the number of parameters of a model. To implement L1 regularization, we place a factorized Laplacian prior on the parameters. The laplacian prior can be defined as follows (Of note, the prior definition is borrowed from the HW2):

$$p(\theta|\lambda) = \prod_{s\in\nu} Lap(\theta_s|\lambda) \prod_{(s,t)\in\varepsilon} Lap(\theta_{st}|\lambda)$$
$$p(\theta|\lambda) = -\frac{\lambda}{2} exp\{-\lambda|\theta|\}$$

To be specific, the λ in above equations are Laplacian parameters.

Training Objective function After placing a Laplacian prior on the parameters, the log-likelihood of MRF structure can be defined as (Here, the normalizer of the prior is ignored):

$$logp(\theta|X,\lambda) = logp(X|\theta) + logp(\theta) = \sum_{s\in\nu}(\theta_s \sum_{l=1}^{1168} x_s^l - \lambda_s|\theta_s|) + \sum_{st\in\varepsilon}(\theta_{st} \sum_{l=1}^{1168} x_s^l x_t^l - \lambda_{st}|\theta_{st}|) - L\Phi(\theta)$$

Gradient Objective After obtaining the training objective function, with standard expression for the derivative of log partition function, the gradient objective with respect to $\theta_s \quad and \quad \theta_{st}$ can be defined as followed (Assuming that $\theta_s \quad and \quad \theta_{st}$ are both **non-negative**):

$$\frac{\partial logp(\theta|X,\lambda)}{\partial\theta_s} = \sum_{l=1}^{1168} x_s^l - \lambda_s - LE_\theta[X_s] \qquad E_\theta[X_s] = \sum_{X'} X_s^{'} p(X^{'}|\theta)$$

$$\frac{\partial logp(\theta|X,\lambda)}{\partial\theta_{st}} = \sum_{l=1}^{1168} x_s^l x_t^l - \lambda_{st} - LE_\theta[X_s X_t] \qquad E_\theta[X_s X_t] = \sum_{X'} X_s^{'} X_t^{'} p(X^{'}|\theta)$$

If $\theta_s \quad and \quad \theta_{st}$ are both **negative**, the gradients are as the following equations.

$$\frac{\partial logp(\theta|X,\lambda)}{\partial\theta_s} = \sum_{l=1}^{1168} x_s^l + \lambda_s - LE_\theta[X_s] \qquad E_\theta[X_s] = \sum_{X'} X_s^{'} p(X^{'}|\theta)$$

$$\frac{\partial logp(\theta|X,\lambda)}{\partial\theta_{st}} = \sum_{l=1}^{1168} x_s^l x_t^l + \lambda_{st} - LE_\theta[X_s X_t] \qquad E_\theta[X_s X_t] = \sum_{X'} X_s^{'} X_t^{'} p(X^{'}|\theta)$$

2.2.3 Implementation Detail

To compute the MAP estimators for the parameters θ, the above gradient functions and the PyProximal optimization package are used. Of note, to simplify, **we do not impose L1 regularization on node parameters, and set the L1 regularization parameter λ_{st} is the same for every edge in one experiment**. A series of experiments are run to see how the L1 regularization parameter λ make a difference to the generated MRF model. Mathematically, in each experiment, λ_s are 0 for all nodes and λ_{st} are same for every edge. Here is a example code to use one function of PyProximal package to obtain estimated θ. After importing **ProximalGradient** from **pyproximal.optimization.primal** in python, the MAP estimators can be computed using the following few lines of code.

```
f =Ising(ssTrain,LTrain,N)
g = Subset_L1(N)
thetaL1[:,ll] = ProximalGradient(f, g, x0=theta0, epsg=lambdaBar, niter=1000)
```

The code to construct MRF models are referenced from the HW2 of this class. In this snippet of code, when looking at ProximalGradient function, f is the log likelihood of a MRF model before impsoing L1 regularization, g represents the node parameters which are free of L1 regularization, x0 is the initialized θ parameters, epsg argument is where to set the λ_{st}, and niter argument is to determine the number of iterations.

3 Experiments

3.1 Data

3.1.1 Data preprocessing

The original dataset (Chen, 2017) consists of hourly measurements of air-quality-related variables for each station. The air-quality related measurements for each region in original dataset contain PM2.5 measurements, PM10 measurements, SO2, NO2 and son on. In this project, we only extracted the PM2.5 values for each region and use PM2.5 as the only indicator for air quality. During the preprocessing phase, we calculated the daily average of PM2.5 values per station and subsequently classified these daily averages into six distinct categories: 0, 1, 2,3,4,5. This classification was

guided by the latest Air Quality Index in US detailed in Figure 3. The categories (0,1,2,3,4,5) respectively represents Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous in Air Quality Index. For example, if a region's daily PM2.5 value falls in the range of 0.0-12.0 ug/m^3, the daily air quality category of this region is good. It is worth mentioning that training/test split ratio is 80:20. The training data has 1168 days of air quality categories of regions in Beijing. The test data has 293 days.

Figure 3: Air quality index

AQI Category	Index Values	Previous Breakpoints (1999 AQI) (μg/m^3, 24-hour average)	Revised Breakpoints (μg/m^3, 24-hour average)
Good	0 - 50	0.0 - 15.0	0.0 – 12.0
Moderate	51 - 100	>15.0 - 40	12.1 – 35.4
Unhealthy for Sensitive Groups	101 – 150	>40 – 65	35.5 – 55.4
Unhealthy	151 – 200	> 65 – 150	55.5 – 150.4
Very Unhealthy	201 – 300	> 150 – 250	150.5 – 250.4
Hazardous	301 – 400	> 250 – 350	250.5 – 350.4
	401 – 500	> 350 – 500	350.5 – 500

3.1.2 Data visualization

After preprocessing the data, we did some EDA analysis to have a sense of the data before crafting the models on data. Figure. 4 illustrates how yearly-average air quality categories of 12 regions (yearly averages are computed from daily averages) change with time. The 12 regions seem to have similar trends on air quality with respect to time. This observation may imply that a fully-connected model is more suitable for the data. However, the Dingling region exhibited lower air pollution compared to the remaining 11 stations.

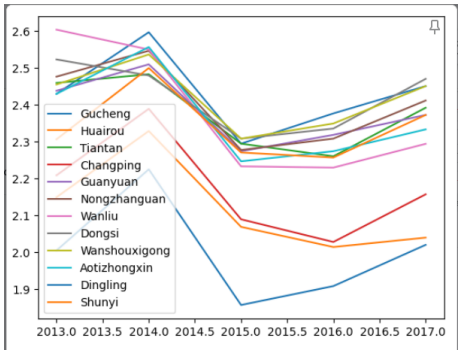
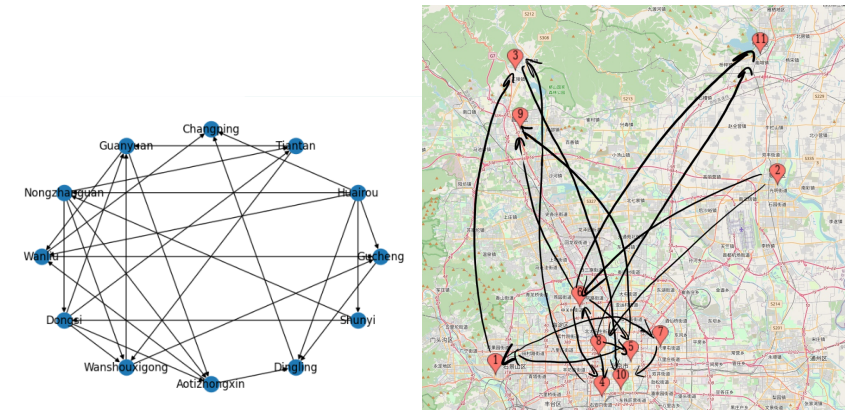


Figure 4: Yearly Average of PM2.5 across 12 stations

3.2 Bayesian Network

3.2.1 Structure Learning

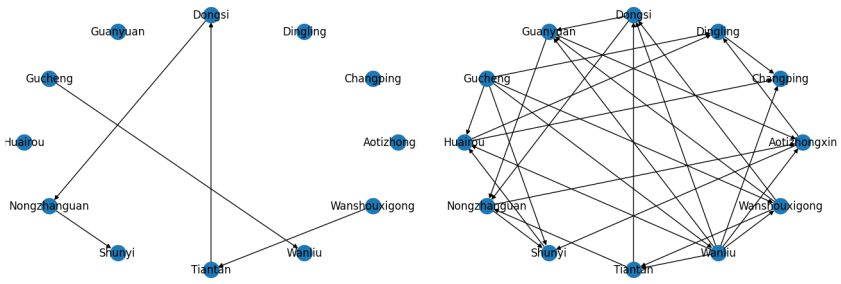
In our initial approach to constructing the BN, we opted to commence the Hill-Climbing search with a fully-disconnected BN. This led us to derive a specific structure for our BN following the search. To delve deeper into the causality, we subsequently represented the stations on a map of Beijing, highlighting a few interconnections. Intriguingly, the causal associations suggested a predominant directionality. The majority of the edges veered upwards and to the right, closely mirroring Beijing's prevailing South-West wind pattern. This observation may hint at a significant role of wind and its direction in the distribution and correlation of pollution across different regions. Based on this understanding, it could be beneficial to initialize the Hill-Climbing search with a graph that already incorporates some edges, reflecting the dominant wind direction.



(a) Learned Structure of a BN (b) Graph on Beijing's map

Figure 5: The learned structure of the graph

In an alternative strategy, we initialized the Hill-Climbing search using a rudimentary graph, which incorporated the directionality of the wind. Interestingly, this graph retained a substantial resemblance to the one generated when initialized with a fully-disconnected BN. Evaluating the K2 metric on both graphs revealed that the modified approach resulted in only a slight enhancement. The K2 score modestly improved from -281905.83 to -281330.29, underscoring that the impact of initial wind directionality in the graph is minimal. Perhaps more edges would be needed for further improvements, although the initial graph should have too many edges which could lead to a more dense structure.



(a) Initial graph with some directionality (b) Learned structure with initial graph

Figure 6: The learned structure of the graph

3.2.2 Inference

After the network structure has been determined, we proceed to deduce the parameters from the training data, utilizing a Bayesian estimator and a K2 prior. In our inference methodology, we employ the PM2.5 measurements from 11 stations as evidence. Consequently, the model strives to deduce the PM2.5 value for the omitted 12th station. The accuracy was fairly consistent across all stations, registering around 0.23. The validation results for both the training and testing datasets demonstrated minimal differences, suggesting that our model doesn't overfit.

3.3 Markov Random Field

3.3.1 Data binarization

The MRF model will be built with Ising model, in which each node (variable) has only two states. To facilitate the following model building, the data are binarized. In Data Preprocessing section, it is known that daily air quality category in each region is classified into six categories (0,1,2,3,4,5). To binarize, the previous categories (0 and 1), indicating Good and Moderate, are combined into a

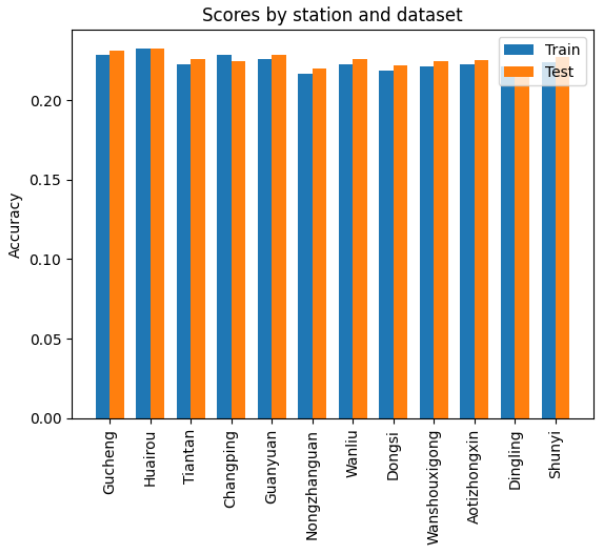


Figure 7: Accuracy for inference

new category (0); the previous categories (2,3,4,5), representing (Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous), are combined into a new category (1).

Fully Disconnected MRF model The log-likelihood of a Fully Disconnected MRF model can be expressed as: $logp(X|\theta) = \sum_{l=1}^{1168}(\sum_{s\in\nu}\theta_s x_s^l - \Phi(\theta))$. Taking the derivative of $logp(X|\theta)$ as zero, we can obtain the ML estimates for θ: $\theta_s = log(\frac{\frac{1}{L}\sum_{l=1}^{1168} x_s}{1-\frac{1}{L}\sum_{l=1}^{1168} x_s})$. By feeding the training data into the formula of ML estimated θ, a Fully Disconnected MRF model can be generated as shown in Figure. 8.

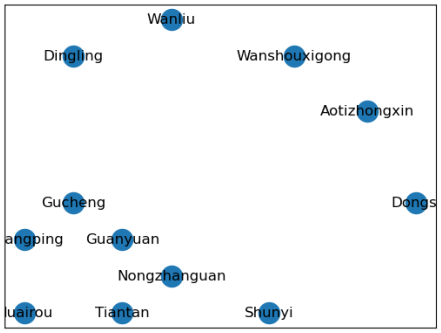


Figure 8: Fully Disconnected MRF model

The Figure. 9 is drawn from the learned parameters θ_s for each region. To be specific, the θ_s can be interpreted as expected log odds ratio for each region. From the Figure. 9, it can be observed that except Dingling region, all other regions tend to have the air quality category below Moderate.

3.3.2 MRF model with L1 regularization

Before exploring the effect of L1 regularization, a **Fully Connected** MRF model is created. Then we tune the strength of L1 regularization to see how the strength influence the generated models. To increase the strength of L1 regularization, we simply increase the λ parameters which are described in the section2.2.2 (Training objective).

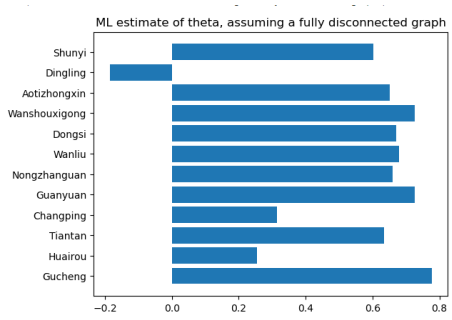


Figure 9: Log odds ratio for each region

Effect of strength of L1 regularization on the edges of MRF models Supposedly, as we increase the strength of L1 regularization, there will be fewer edges in generated MRF models. However, in Figure. 10 and Figure. 11, it can be seen that initially, the increment of L1 regularization cannot make the generated model less dense or remove any edge in MRF models; when the lambda reaches a certain threshold around 450, the fully connected MRF models suddenly turn to be fully disconnected MRF models. It is a interesting phenomenon. The reasons behind this phenomenon may be the 12 regions in Beijing are inherently closely related, because the 12 regions are geographically very close and in the same city. The inherent relationship across 12 regions can also be spotted from the above Figure.4, which shows that the yearly average air quality of 12 regions have a similar trend on time.

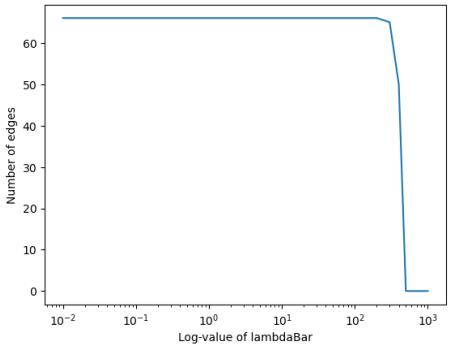
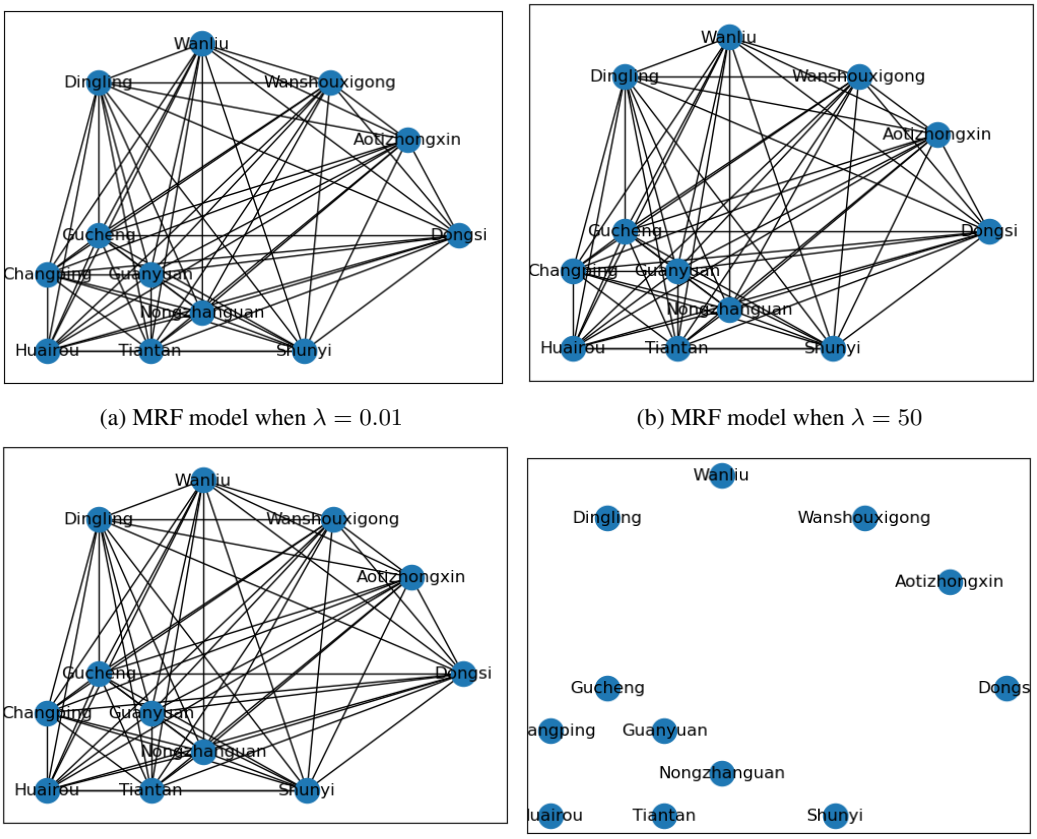


Figure 10: Log odds ratio for each region

Effect of strength of L1 regularization on the log-likelihood of models After creating the models under different L1 regularization, we also compare the log-likelihood of the models. In Figure. 12, it can be observed **fully connected** models have highest log likelihood compared with the models in which some edges are removed. When part of edges in fully connected models are removed, the log likelihood are dropping. Meantime, we can see that the log likelihood of the models have a sharp increase again when the models become **fully disconnected**, although the log likelihood of the fully disconnected models are still much lower compared with fully connected models. One potential reason for why fully connected and fully disconnected models tend to have relatively higher likelihood is still that 12 regions are very geographically close and there is inherently close relationship between 12 regions.

3.3.3 Evaluation on test data

We assume the MRF models with higher log-likelihood are precise to characterize the dependencies across 12 regions. We then select four best learned models from training data with relatively highest log likelihood to evaluate their prediction accuracy on test data. To compute prediction accuracy, we let a MRF model use other eleven regions' air quality category to predict the air quality category of a target region. The accuray can be defined as the number of days with correct prediction over the total number of days in test data. In following Figure. 13, it can be seen that the four models' prediction



(a) MRF model when $\lambda = 0.01$

(b) MRF model when $\lambda = 50$

(c) MRF model when $\lambda = 300$

(d) MRF model when $\lambda = 500$

Figure 11: The MRF models under different L1 regularization

accuracy of four models are same with respect to each region. The exact equivalence of four models' prediction accuracy is because 4 models are generated under very close L1 regularization, the λ for four models are 0.01, 0.05, 0.1, and 0.5. Averaging prediction accuracy over regions, four models' prediction accuracy on test data is 0.905.

4 Conclusion and future works

4.1 Bayesian network

We used Bayesian networks for structure learning on Beijing's pollution data, unearthing patterns reminiscent of local wind trends. While initializing the Hill-Climbing algorithm with the dominant wind direction accelerated convergence, it did not drastically alter outcomes.

The prediction accuracy was suboptimal, potentially due to the ordinal nature of our data and our focus on exact matches during evaluation. This approach may be considered restrictive, as predicting class 1 for a true class 0 is markedly closer than predicting class 5.

Including meteorological features like precipitation and wind speed could potentially improve analysis. Despite our inability to explore these variables due to time constraints, they provide promising directions for future studies.

4.2 Markov Random Field

After building MRF models on the data, it can be found L1 regularization can play a significant role in the structure of MRF models. Although in this project, we cannot see a smooth decrease of edges in MRF models with stricter L1 regularization as usual, we can still see that MRF models turn from

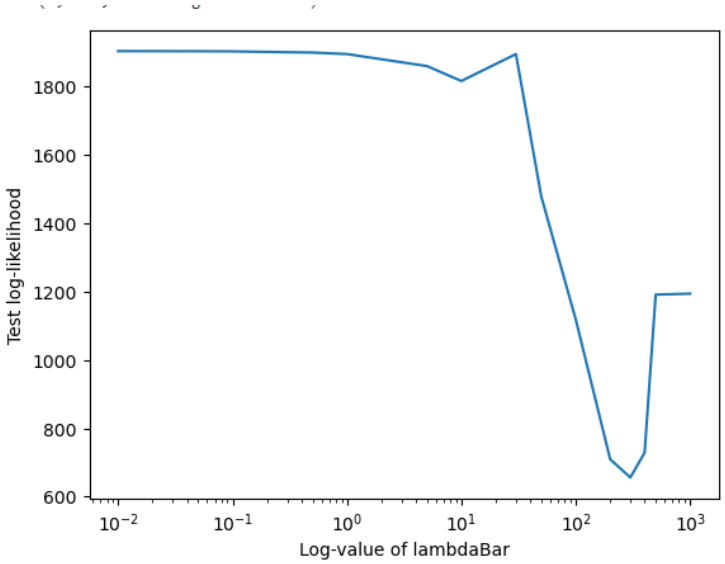


Figure 12: Log-likelihood VS L1 regularization strength

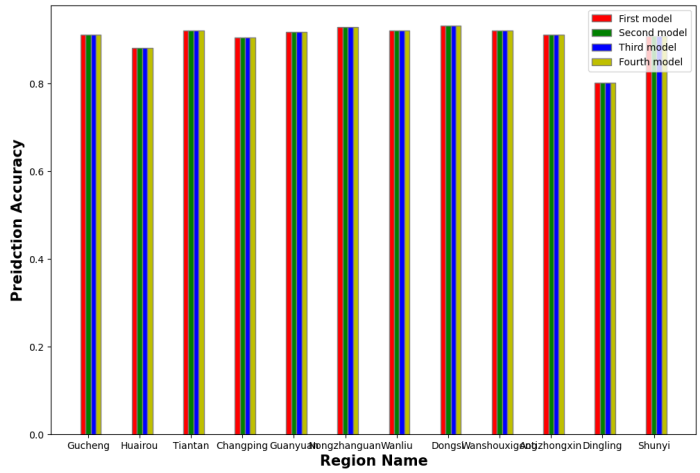


Figure 13: Prediction Accuracy of the models with highest likelihood

fully connected models to fully disconnected models when turning up L1 regularization to a certain value. During tuning L1 regularization, we found the models with fully connected structure have higher likelihood. If L1 regularization is not very strong (λ less than 400), most learned MRF models are approximate to fully connected network.The reasons behind these may arise from the dataset used in this project. The dataset in this project are air quality condition of 12 regions in a same city, maybe the air quality in 12 regions are essentially almost identical because of geographical closeness. If the dataset can contain air quality information of some regions which are relatively distant from each other but are still connected, by incorporating these regions into MRF models, we may observe the MRF models with different density under different L1 regularization. From this project, one lesson is that the selection of dataset is important. The inherent pattern in the dataset will hugely influence the generated models' structure.

When evaluating the learned MRF model with prediction accuracy, the MRF models with high log likelihood also have high accuracy scores on test data. High prediction accuracy indicate that our models capture the patterns in the dataset.

References

Dan Murtaugh. Beijing choked by worst air pollution in more than a year, 2022. URL https://www.bloomberg.com/news/articles/2022-12-12/beijing-choked-by-worst-air-pollution-in-more-than-a-year#xj4y7vzkg.

Antonio S Cofino, Rafael Cano, Carmen Sordo, and Jose M Gutierrez. Bayesian networks for probabilistic weather prediction. In *15th Eureopean Conference on Artificial Intelligence (ECAI)*. Citeseer, 2002.

Aye Nandar. Bayesian network probability model for weather prediction. In *2009 International Conference on the Current Trends in Information Technology (CTIT)*, pages 1–5. IEEE, 2009.

Serdar Dindar, Sakdirat Kaewunruen, Min An, and Joseph M Sussman. Bayesian network-based probability analysis of train derailments caused by various extreme weather patterns on railway turnouts. *Safety science*, 110:20–30, 2018.

Christian Borgelt and Rudolf Kruse. An empirical investigation of the k2 metric. In Salem Benferhat and Philippe Besnard, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 240–251, Berlin, Heidelberg, 2001. Springer. ISBN 978-3-540-44652-1.

Song Chen. Beijing PM2.5 Data. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5JS49.