Learning a probabalistic model of rainfall using graphical models

Byoungkoo Lee
Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
byounko@andrew.cmu.edu

Jacob Joseph
Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
jmjoseph@andrew.cmu.edu

Abstract

We present an analysis of historical precipitation data for the United States' Pacific Northwest, measured for the years 1949-1994 on a grid of approximately 50km resolution. We have implemented a Bayesian network with nodes representing individual geographic grid regions. Directed, weighted edges represent dependence relationships between regions. Using a modified K-2 learning algorithm, we build a heuristically optimal Bayesian network. We examine degree of dependence between regions, the predictive capacity of a minimal set of measurements, and evaluate the utility of additional strategically selected measurements in enhancing local predictions.

1 Introduction

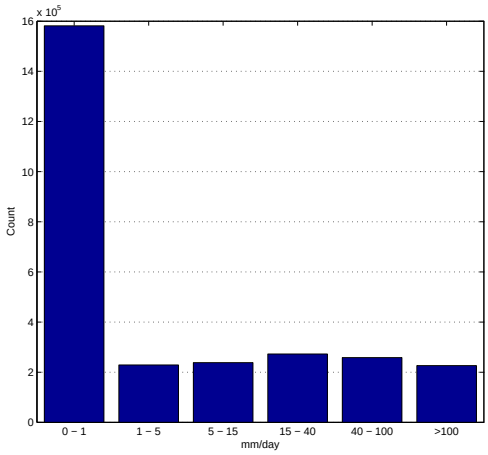
Although weather prediction is essential to many of our social and economic processes, accurate prediction remains an open field of research. On the most simplistic level, weather derives from a variety of interdependent physical factors, including wind speed, air pressure, temperature, ocean currents, and local topology. Meteorologists typically rely upon numerical atmospheric circulation models (ACMs) to predict local and global weather at short and long time scales. These models are most effective at low resolution, predicting large-scale events [1].

An orthogonal approach to weather prediction involves statistical models constructed from local historical data. Such models are typically designed to represent local effects. Many machine learning techniques such as Markov chains, auto-regressive models, and neural networks have been used with limited success. In particular, these models fail to represent spatial and temporal dependencies between neighboring locales [1].

In this study, we examine the use of Bayesian networks to better capture regional dependencies in the limited context of precipitation prediction. We are particularly interested in determining a minimal set of measurement sites sufficient to quantitatively predict local rainfall. Central to these goals, we exploit the interdependence between geographically disparate measurements to evaluate the utility of each existing measurement site and potential new sites.

From the given historical data, we construct a high-resolution ($<=$50km grid) probabilistic

Figure 1: **Histogram of Rainfall (mm)**



model of rainfall throughout the Pacific Northwest. We are particularly interested determining a minimal set of measurement sites sufficient to quantitatively predict local rainfall. Central to these goals, we exploit the interdependence between measurements at distinct stations and geographic regions to evaluate the utility of each data source.

2 Data

We have been provided precipitation data derived from a number of measurement sites throughout the United States' Pacific Northwest [2]. This data is formatted to a grid of 17 discrete latitudes and 16 discrete longitudes. The actual measurement stations within each grid cell have been consolidated. Several cells have no measurement sites. For each geographical area, a daily measurement of rainfall is provided for the years 1949-1994, totaling 16801 daily measurements. Due to the nature of the data collection, some locations do not include daily measurements over the period considered. The few grid points, or nodes, with incomplete measurements over the full time series have been omitted for simplicity. All analyses have been performed using the 167 nodes with complete data series.

Data pre-processing consisted of conversion from the provided netCDF format to a native three-dimensional Matlab array more amenable to analysis without additional Matlab interfaces. Several such Matlab-netCDF interfaces are available, though none proved usable with the particular Matlab environment available to us. An indirect approach was accomplished by first transforming the netCDF format to ASCII using native libraries, and finally reconstructing a multidimensional Matlab array.

2.1 Discretization

Daily rainfall measurements are supplied as continuous values of millimeters per day. To facilitate construction of a discrete Bayesian network, we opted to to discretize rainfall to six categories, corresponding to 0-1 ('no rain'), 1-5, 5-15, 15-40, 40-100, and $>$100 mm/day, respectively. This approach as been used previously to represent light, medium, and heavy rain [1]. The histogram within Figure 1 illustrates the number of measurements observed within each category. We sought to minimize data skew by empirically selecting thresholds to represent equal-sized populations within each category. A roughly exponen-

Figure 2: **Modified K-2 Algorithm**

```
Input: Quantized data of n nodes, an ordering of n nodes,
       an ordering of neighbors for each node, max_parents
Output:Adjacency matrix representing all directed edges in the
       network

For i = 1 to n
  parent_i = []; #Initial condition: no parent node for any node
  P_old = f(i, parent_i); #Probability of data (i node) given parent_i
  Gonext = true;
  While Gonext & size(parent_i) < max_num
    P_new = f(i, parent_i, another parent_i); # choose from neighbors
    If P_new > P_old
      P_old = P_new;
      parent_i = parent_i + another parent_i;
    else Gonext = false;
  end
  Save parent nodes for node i; # in adjacent matrix
end
return adjacency matrix
```

tial decrease from 0mm/day in the number of measurements is observed, resulting in increasing category bin widths. Note that all values within the 0-1 category are exactly 0 and are thus insensitive to threshold selection. Note that Euclidean distances and correlation were calculated with the original continuous data series.

3 Methods

3.1 Bayes Network Construction

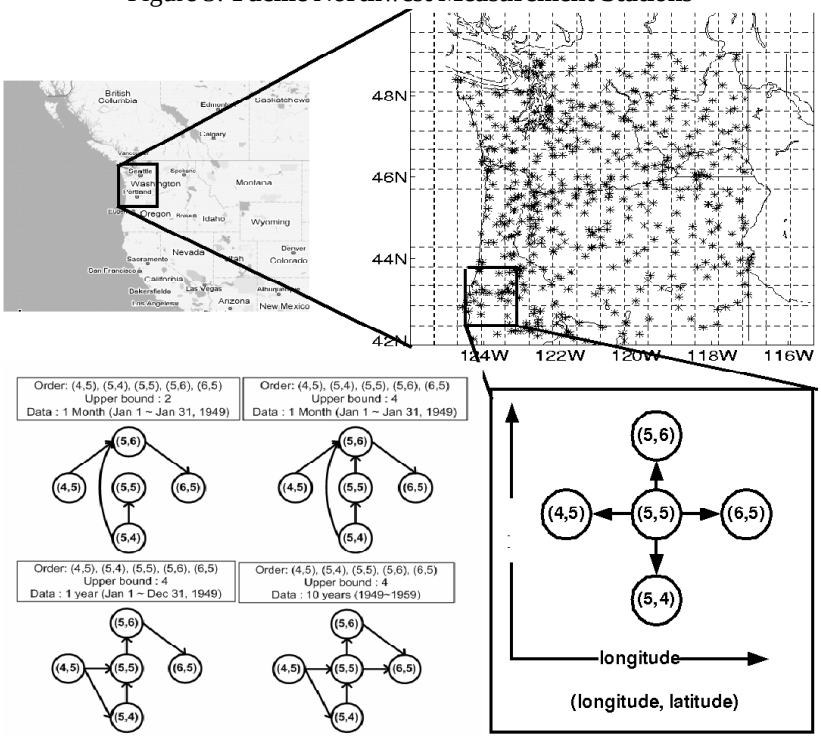
As each node in a Bayes network may conditioned upon any other node in the network, exhaustively learning an optimal network structure for all but the smallest networks is computationally intractable. Indeed, this problem is NP-hard. As such, a number of heuristics are commonly used to approximate a globally optimal DAG structure. These include the Metropolis-Hastings Markov Chain Monte Carlo (MCMC) method to sample the DAG space, hill climbing methods to explore node neighbors incrementally, active structure learning[3], and structural EM [6]. We utilized the K-2 algorithm[7] due to its ease of implementation and suitability for subsequent modification.

An ideal network structure maximizes the probability of the network given the observed data, $\text{argmax}_{net}\text{P(net}|\text{data)}$. Using Bayes' rule and a constant P(data), $\text{argmax}_{net}\text{P(net}|\text{data)} = \text{argmax}_{net}\ \frac{P(net,data)}{P(data)} = \text{argmax}_{net}\ \text{P(net, data)}$. Structure learning algorithms score potential networks based upon this latter property.

The standard K-2 algorithm is a greedy algorithm that iteratively selects parents for each node independent of other all other nodes. Beginning with an empty set of parents at each node, the method incrementally adds any single parent which increases the node's overall score. This is repeated until no single additional parent can increase the score. The method does require a fixed node ordering to avoid cycles.

We have examined two efficiency improvements of the K-2 algorithm to facilitate practical computation of our large, 167 node network structure. First, while the standard K-2 im-

Figure 3: **Pacific Northwest Measurement Stations**



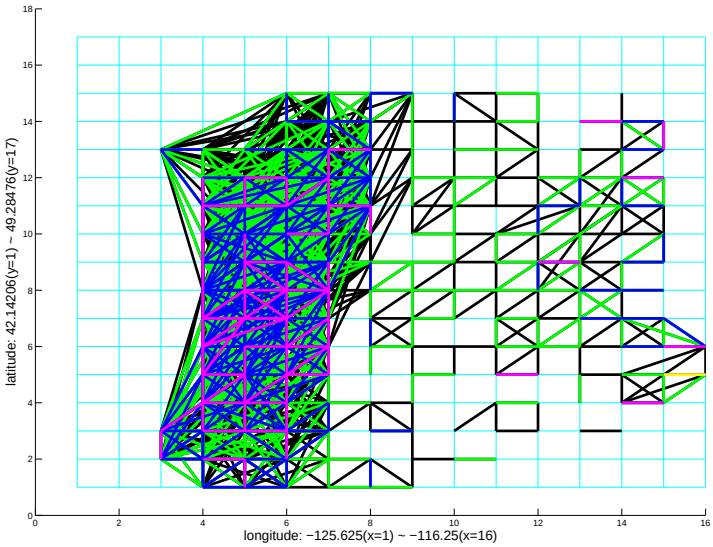
plementation iteratively adds all beneficial parents from the set of previously visited nodes, we have set an upper limit on the number of possible parents. Secondly, we have limited the set of parents considered to significantly less than the complete node ordering. We have modeled this after the concept of the *Local K-2* algorithm, described in [1]. The *Local K-2* algorithm was used to consider only a node's *k*-nearest neighbors, measured by correlation. In addition to the

We have examined the efficacy of both correlation and and euclidean distance in selecting potential parents. Note that the pairwise euclidean distances and correlation were calculated with the original, continuous data series.

4 Results and Discussion

As an initial evaluation of our approach, we selected a subset of five grid regions. We then calculated a topological structure of these stations by K-2 algorithm. As a naive approach to node ordering for this heuristic, we presented nodes in a Southwest to Northeast order, consistent with global wind patterns for the area. Figure 3 indicates the grid subregion selected. Note that (*) indicates a physical measurement station from which grid cell values have been derived. Several example learned topologies are shown for durations of 1 month, 1 year, 10 years, and two selections of the maximum number of parents per node. These serve to illustrate the sensitivity to parameter selection.

Figure 4: **Correlation Between Measurement Nodes**



Network Structure

Figure 4 illustrates the correlation between measurement sites. The correlation coefficient between sites was calculated using all 16801 continuous measurements from each. To compactly represent the greatest correlation coefficients, we draw a colored line between two sites as follows: yellow (corr $>$ 0.95), magenta (corr $>$ 0.90), blue (corr $>$ 0.85), green (corr $>$ 0.80), and black (corr $>$ 0.75). The densely connected, and highly correlated, Western region is of meteorological significance in that the Cascade mountain range runs vertically at approximately the observed longitude. This range separates a relatively wet, coastal climate from the comparatively more arid Eastern region.

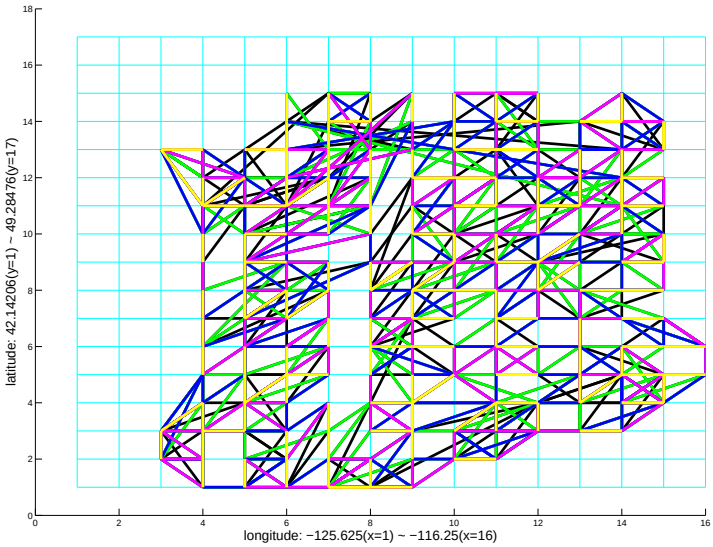
Measurement sites exhibit greatest correlation with their immediate geographic neighbors. In the West, knowledge of precipitation anywhere throughout this highly connected component facilitates accurate inference. By contrast, sites in Eastern longitudes are less similar to both immediate as well as distant neighbors. This may reflect an inability of our model to perform well in arid regions or that local climate effects are more significant. Additional measurement sites would be best placed within these sparsely correlated regions to enhance the predictive power of the resulting Bayesian network model.

We have also considered euclidean distance of the time-series data to aid in selecting appropriate parents of a node. Figure 5 shows a ranking of distance to neighbors in this space. From closest distance, the coloring is yellow, magenta, blue, green, and black. This plot indicates the best possible selection of neighbors and clearly demonstrates that a site's immediate geographic neighbors are the best predictors.

Bayesian Inference

The ten nearest neighbors of each node, in euclidean distance, were used to learn the graph structure from our complete data set. This graph structure was then used as a basis to train

Figure 5: **Euclidean Distance Rank**



a Bayesian network using conditional probabilities derived from the observed frequencies. We evaluated this Bayesian network by several methods. First, we examined subsets of nodes known to be highly correlated. As an example, Table 1 illustrates the sensitivity of the measurement site at longitude 6, latitude 15 (6,15) to the value of its neighboring parents, (6,14) and (7,15).

This is representative of all highly correlated nodes. Interactions are also present across greater geographic distances within correlated components. In general, the conditional probability tables in this propagation contain fewer zero probabilities (and thus are not shown).

The second major approach utilized to examine the quality of the Bayesian network was to examine the data likelihood. Figure 6 shows the log-likelihood of the data for each of the original 16801 data samples. Despite construction of the Bayesian network from the complete data set, the data at many time-points exhibit surprisingly low log-likelihoods. We expect this may be due to sub-optimal data partitioning, possibly with respect to seasonal changes. Time permitting, we aim to further examine any periodicity in time of this log-likelihood to establish partitions of the data. Initial data cross-validation results do not indicate significant network structure changes among random data samples. Additionally, the log-likelihoods within a test set are very similar, further suggesting that our method performs similarly across the data set.

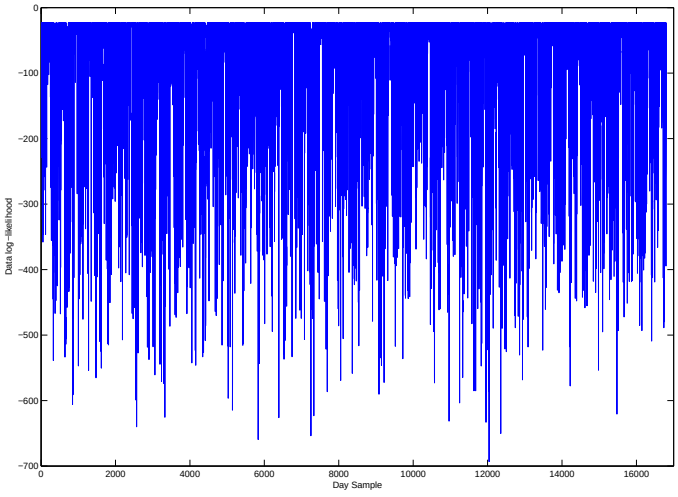
Table 1: **Marginal and Conditional Probabilities of Site (6,15)**

value	P((lat6,lon15) = value)
1	0.6060
2	0.0554
3	0.0814
4	0.1027
5	0.1012
6	0.0534

(6,15)	(7,15)	(6,14)	P((6,14 | (6,15), (7,15))
1	1	1	0.9882
1	1	2	0.0102
1	1	3	0.0016
1	1	4	0.0000
1	1	5	0.0000
1	1	6	0.0000
6	6	1	0.0000
6	6	2	0.0000
6	6	3	0.0044
6	6	4	0.0087
6	6	5	0.1004
6	6	6	0.8865

All omitted values have zero probability.

Figure 6: **Daily Sample Log-likelihood**



5 Conclusions

Using our modified K-2 algorithm, we have been able to rapidly build a Bayesian network representative of rainfall in the US Pacific Northwest. We also consider dependence between measurement sites using correlation coefficients as well as distances of time series. Based on the pattern of correlation coefficients, we can propose additional measurement sites to get a higher prediction on the area. From the figure of correlation and Bayesian network, we can conclude that Bayesian network appropriately represents the dependence of measurement sites. Future directions include additional evaluation and comparison of distinct network structures constructed from subsets of measurement sites. While we sought to examine the limited context of inference within single time-steps, the consideration of temporal interactions, possibly using a dynamic Bayesian network, may yield increased accuracy and facilitate forecasting.

References

[1] Rafael Cano, Carmen Sordo & Jose M. Gutierrez (2004). *Applications of Bayesian Networks in Meteorology.* Advances in Bayesian Networks, Springer, 309-327.

[2] Martin Widmann and Christopher Bretherton. *"50" km resolution daily precipitation for the Pacific Northwest, 1949-94.* http://www.jisao.washington.edu/data_sets/widmann

[3] Murphy K.P. (2001). *The Bayes net toolbox for Matlab.* Computing Science and Statistics 33. http://bnt.sourceforge.net

[4] Martin Widmann & Christopher S. Bretherton (1998). *Validation of mesoscale precipitation in the NCEP reanalysis using a new grid-cell data set for the northwestern United States.* Journal of Climate.

[5] Antonio S. Cofino, Rafael Cano, Carmen Sordo & Jose M. Gutierrez (2002). *Bayesian Networks for Probabalistic Weather Prediction.* ECIA 2002, Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press, 695-700.

[6] Friedman et al. (1997) *Bayesian Network Classifiers.* Machine Learning, Springer, 131-163.

[7] Cooper and Herskovitz (1992) *A Bayesian method for the induction of probabilistic networks from data.* Machine Learning, Springer, 309-347.