Rain

Matthew Wiens, Kelly Kung

December 20, 2017

Abstract

Predicting precipitation is a key problem for residents of the Pacific Northwest, from it's impact on family vacations to landslides. A noteworthy feature of the area is distinct chances of precipitation between the summer and winter. In this report, we propose a model for differentiating the two seasons and understanding the chance of precipitation in each, across different climatic areas of the state. Such a model can be used to predict optimal times to travel from a rainy region to a dry region. We adopt a Bayesian methodology and discuss predictive results within the Bayesian framework.

1 Introduction

Residents and visitors to the Pacific Northwest (Oregon and Washington) notice a distinct pattern year over year: most days during the winter are rainy, and then they feel that there a distinct point during the spring or early summer when the weather becomes dry most days. Similarly, during early fall the nice weather switches back to rain. However, the timing of the switch varies throughout the area, with the prominent Cascade Mountains being a key factor. There are many resorts on the east side of the Cascade Range that cater to residents of the west side looking to escape the rain.

Widmann and Bretherton developed a methodology to control for local variation in topography and precipitation data and generated a dataset of estimated precipitation data over 46 years in a 50km by 50km grid. In comparison, general models of atmospheric weather operate on the order of hundreds of kilometers, so there significant room to improve the spatial resolution of weather models. In particular, local topographic features are not captured by general model, which is a limitation in regions like the Pacific Northwest. The temporal and spatial correlation is also highly dependent on the topography and difficult to model, and therefore there has been a focus on parametrizing models and reducing the dimensionality.

In this report, we propose a Bayesian approach for two reasons: first, to capture prior beliefs from living in the area of one author, and second to be able to discuss distributions around complicated model parameters and probabilistic beliefs about the state of weather on a specific day of the year. Bayesian approaches for daily precipitation data are not new, for example see [Olson and Kleiber].

2 Background

2.1 Dataset

To analyze the precipitation in the Pacific Northwest, we use the same dataset as Widmann and Bretherton. This data contains precipitation logs (measured in millimeters) from 1949 - 1994 that was collected form the National Weather Service, which was then corrected by the National Climatic Data Center[1][2]. The

[1] The data was corrected during the 'Validated Historical Daily Data' project, which aimed to verify the data collected.

[2] T. Reek, S. Doty, and T. Owen. A Deterministic Approach to the Validation of Historical Daily Temperature and Precipitation Data from the Cooperative Network. In Bulletin of the American Meteorological Society 73, no. 6 (1992): 753-62.

dataset is a 3-dimensional array with a 2-dimensional grid-cell that consists of 16 longitude subsections and 17 latitude subsections, where each cross-section is approximately 0.48° (latitude) by 0.62° (longitude). The third dimension contains temporal data (16,801 days), which means we have approximately 4 million data points. However, note that there are approximately 32,000 missing values which correspond to areas with few functional weather stations, such as the ocean and some areas in the southeast. Using this dataset, we then transformed and created several variables in order to obtain a model to analyze the precipitation.

2.2 Grouping of Areas

The Pacific Northwest has a varied climate, from temperate rainforests on the Pacific coast to desert in Southeastern Oregon, which has a major impact on the seasons. Therefore, each grid cell is classified into one of nine climatic zones as follows:

1. Washington Coastline (Temperate Rainforest)
2. Oregon Coastline
3. Western Washington / Seattle Metropolitan Area
4. Portland Metropolitan Area / Willamette Valley / Southwestern Oregon
5. Cascade Range
6. Columbia Plateau
7. Selkirk Mountains
8. Blue Mountains
9. Eastern Oregon / High Desert / Great Basin

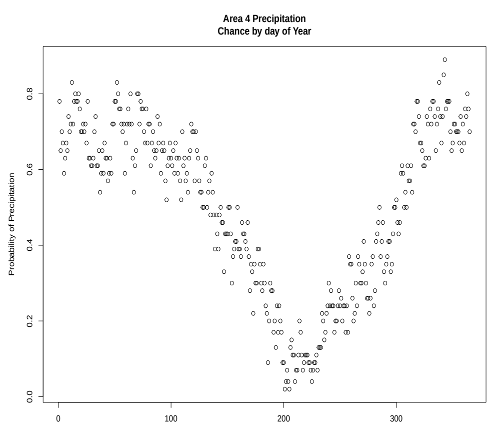

Figure 1: Western Oregon precipitation chance over the year

These zones were decided by considering average precipitation over the time period of the data and the topographic features of the region. The rainfall for the region was aggregated by considering if there was measurable rainfall in each grid by day, and then for each day if the majority of grid cells reported rain, then a value of *rain* was assigned to the area, else it was *dry*

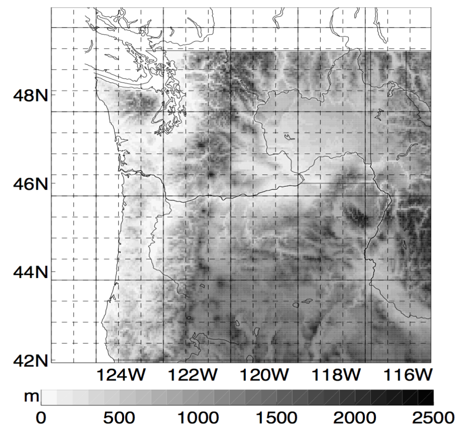

Figure 2: Topography of the region

2.3 Indicators for Dry and Wet Season

Using the grouping of areas as mentioned in Section 2.2, we then proceeded to create indicators of when the dry season and wet season starts. To do this, we first created indicators of precipitation for each day in each area, where a 1 indicates that more than 0.1 mm of precipitation occurred. We then aggregated the indicators in each of the 9 areas for each day and determined that that there was precipitation in the area for that day if a majority of the cells in the area had more than 0.1 mm of precipitation. The areas with precipitation were deemed to be *wet* and the areas without precipitation were deemed to be *dry*. Afterwards, we created an algorithm that chooses indicators of when the dry season starts (X_{min}) and when the wet season starts (X_{max}) such that it maximizes the number of *dry* days in the dry season and maximizes the number of *wet* days in the wet season. For example, we expected the X_{min} days to be approximately 170 and the X_{max} days to be approximately 260 because these dates correspond to the beginning of summer and the beginning of autumn. Using this algorithm, we found the start dates for the dry and wet seasons for the 46 years. [3]

3 Model

3.1 Model 1

Our first model is motivated by the goal of determining when the dry and wet seasons start. To do so, we use a model with a semi-conjugate multivariate normal prior and multivariate normal likelihood. We chose this model because it has a closed form and we are able to sample from the posterior distribution using a Gibbs Sampler on the full conditionals. Furthermore, upon initial exploration of the data, we see that the start dates of dry and wet seasons are approximately normal as seen in Figure XXX. (skeptical about the other plot...?) Thus, we proceeded with the multivariate normal model with confidence that it represents the information we have. We define the model using the following semi-conjugate prior distributions.

$$\theta \sim MVN(\mu_0, \Lambda_0)$$
$$\Sigma \sim Inverse - Wishart_{\nu_0}(\Sigma_0^{-1})$$

[3]Note that when creating the indicators for years, we did not take into account of leap years, which resulted in data points that spilled over to the next year. We just take the years in which we have a year's worth of data for.

To determine the prior parameters, we set $\mu_0 = \begin{bmatrix} X_{min} \\ X_{max} \end{bmatrix} = \begin{bmatrix} 170 \\ 260 \end{bmatrix}$ and $\Lambda_0 = \begin{bmatrix} 30^2 & 0 \\ 0 & 30^2 \end{bmatrix}$ for all 9 area groups. The prior means were chosen according to when the summer and autumn season starts and the prior standard deviations of the $\theta's$ were chosen to represent a spread of one month. We chose the Σ_0 values, shown in to equal the *variance - covariance* matrices for each area group from the data. We chose the prior

1	**2**	**3**	**4**	**5**
$\begin{bmatrix} 1033.00 & -254.79 \\ -254.79 & 562.52 \end{bmatrix}$	$\begin{bmatrix} 652.62 & 32.22 \\ 32.22 & 307.61 \end{bmatrix}$	$\begin{bmatrix} 1103.09 & -35.55 \\ -35.55 & 860.19 \end{bmatrix}$	$\begin{bmatrix} 661.74 & 17.54 \\ 17.54 & 197.88 \end{bmatrix}$	$\begin{bmatrix} 943.12 & 67.83 \\ 67.83 & 524.46 \end{bmatrix}$

6	**7**	**8**	**9**
$\begin{bmatrix} 162.92 & -15.51 \\ -15.51 & 108.72 \end{bmatrix}$	$\begin{bmatrix} 778.02 & -17.29 \\ -17.29 & 99.43 \end{bmatrix}$	$\begin{bmatrix} 1097.19 & 58.49 \\ 58.49 & 76.20 \end{bmatrix}$	$\begin{bmatrix} 288.11 & -15.03 \\ -15.03 & 12.34 \end{bmatrix}$

Table 1: Table of prior Σ_0 values for each area group

In order to conduct Gibbs Sampling, we need full conditionals of θ and Σ.

$$\theta|\Sigma, y \sim MVN(\mu_n(\Sigma), \Lambda_n(\Sigma)))$$
$$\Sigma|\theta, y \sim Inverse - Wishart_{\nu_n}(\Sigma_n^{-1}(\theta))$$

3.2 Model 2

We further developed our model by considering the probability of rain on each day for the summer (dry) and winter (rainy) seasons. This is motivated by the idea of trying to plan outdoor activity in each area of the Pacific Northwest and deciding if there is value in traveling to a different area to escape rain. Therefore, the following model is proposed:

- The data is binary data representing if it rained or not for each day of the year,
- π_w and π_d are the probability of rain during the wet season and dry season, respectively.
- θ_1 and θ_2 are the cutoff points for the wet and dry seasons. The wet season runs from January 1st until the day before θ_1 and from θ_2 to December 31st, and the dry season runs from θ_1 until the day before θ_2, and the natural restriction is imposed that $\theta_1 < \theta_2$
- Let y_j be the data on day j. Then $p(y_j) = \pi_w 1_{j<\theta_1} + \pi_d 1_{\theta_1 \leq j < \theta_2} + \pi_w 1_{j \geq \theta_2}$

4 Analysis

Using the semi conjugate normal model described in section 3.1, we used a Gibbs sampler to generate samples from the posterior distributions.

In figure 3 this model is used to explore how one could escape the rainy season in the major metropolitan areas of Seattle, Washington (region 3) and Portland, Oregon (region 4). For example, resort operators in the drier climates directly east of these cities are most desirable when they have dry weather and the cities have rainy weather. We find that there is longer period in the springtime when going to the east side of the Cascades is desirable, and a narrow window during the fall. However, by noting the maximum probability is substantially less than 1 in each peak, we note that the range of this time period is variable year to year, and so making exact future plans would be difficult.

Region	Start Date Mean	Start Date Variance	End Date Mean	End Date Variance
1	157	22	275	12
2	145	14	286	7
3	159	24	272	19
4	137	15	291	5
5	160	20	276	11
6	110	5	297	3
7	117	18	297	3
8	134	24	298	2
9	114	8	301	1

Table 2: Posterior Mean and Variance of mean parameter

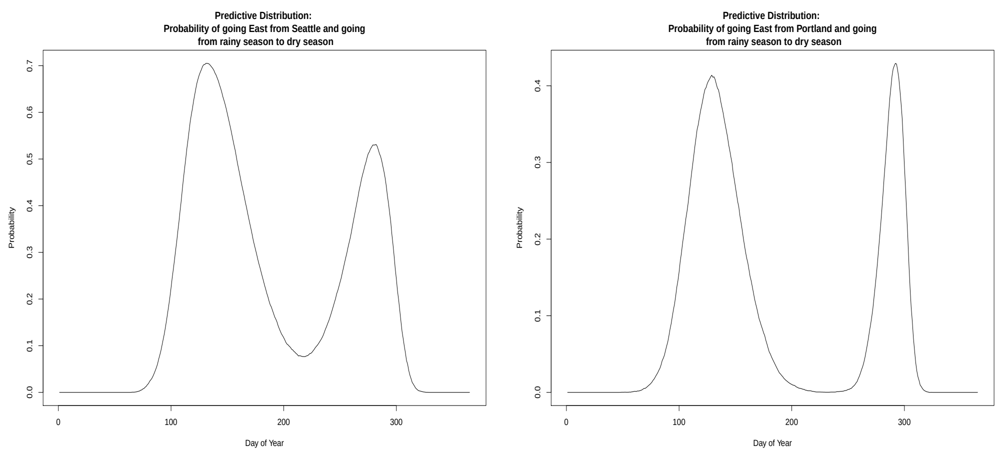

Figure 3: On each day of the year, the probability of escaping the rain by going east for the two major metropolitan areas

5 Discussion

6 Conclusion