Rain

Matthew Wiens, Kelly Kung

December 20, 2017

Abstract

Predicting precipitation is a key problem for residents of the Pacific Northwest, from its 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 are rainy during the winter, and then they feel that there a distinct point during the spring or early summer when the weather becomes mostly dry. 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[1] 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 is significant room to improve the spatial resolution of weather models. In particular, local topographic features are not captured by general models, which is a limitation in regions like the Pacific Northwest. The temporal and spatial correlation is also highly dependent on the topography and are 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].... With a Bayesian approach, we hope to model (1) the beginning of dry and wet seasons and (2) the probability of precipitation in the different areas. These models will provide us with tools in order to further analyze and understand the precipitation patterns in the Northwest.

This report is organized as follows. In Section 2, we refine our goals with our analysis, which motivates the creation of several variables in our dataset, which is also outlined in the same section. In Section 3, we discuss the two models that we used in order to analyze the data. We then begin our analysis of the dataset in Section 4, and in Section 6, we conclude our report with summaries of our analysis as well as potential future directions.

[1] M. Widmann and C. S. Bretherton. Validation of Mesoscale Precipitation in the NCEP Reanalysis Using a New Gridcell Dataset for the Northwestern United States. In Journal of Climate, 13(11), 1936-1950. 2000.

2 Background

2.1 Goals-change name later

In their analysis, Widmann and Bretherton found that there was some variability in precipitation levels in the different topographical areas in the Northwestern United States. For this report, we further investigate the variability in the different areas in order to build upon their results. In particular, we are interested in analysis regarding when the dry and wet seasons start and the probability of precipitation on a given day. By using a Bayesian approach, we are even able to calculate posterior probabilities in order to answer more questions. For example, which area has the longest summer? Which area is most likely to have the latest winter? What is the chance that someone can escape the bad weather by traveling to another part of the are? With our models and analysis, we may be able to better understand the precipitation patterns in the different topographical areas of the Pacific Northwest.

2.2 Precipitation Dataset

To analyze the precipitation in the Pacific Northwest, we use the same dataset as Widmann and Bretherton. This data contains 46 [2] years of precipitation logs (measured in millimeters) from 1949 - 1994 that was collected form the National Weather Service and was corrected by the National Climatic Data Center[3][4]. The dataset is a 3-dimensional array with a 2-dimensional 16 (latitude) by 17 (longitude) grid cell[5] for each of the 16,801 days. However, 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.

2.3 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

[2] 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.

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

[4] 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: 753-62. 1992.

[5] Each cell represents an area approximately 50km by 50km.

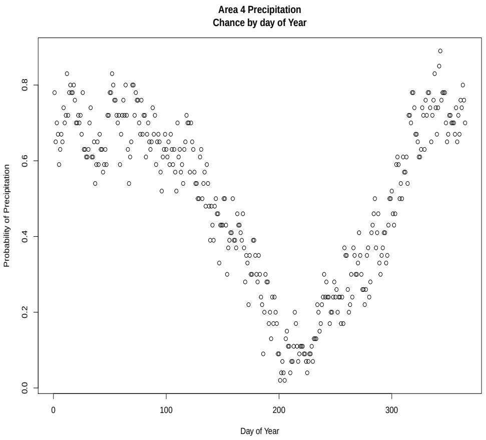

Figure 1: Western Oregon precipitation chance over the year

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 precipitation (more than 0.1 mm) in each grid by day, and then for each day if the majority of grid cells reported precipitation, then a value of *rain* was assigned to the area, else it was *dry*.

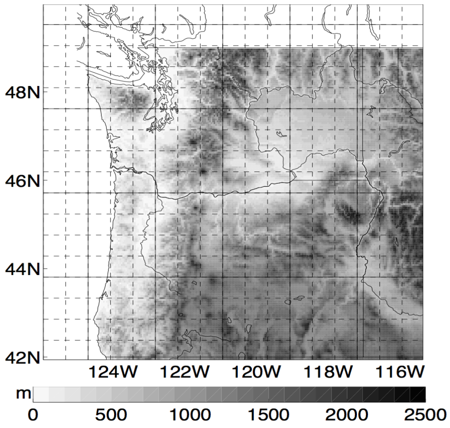

Figure 2: Topography of the region

2.4 Indicators for Dry and Wet Season

Using the area groups and daily precipitation indicators, mentioned in Section 2.3, we then proceeded to create indicators of when the dry season and wet season starts. To do this, we create an algorithm that chooses indicators of start days for the dry season (X_{min}) and wet season (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 expect X_{min} to be approximately 170 and X_{max} to be approximately 260 because these dates correspond to the beginning of summer and the beginning of autumn. Using this algorithm, we find the start dates for the dry and wet seasons for the 46 years.

3 Modeling

3.1 Multivariate Normal Model

Our first model is motivated by the goal of determining when the dry and wet seasons start. We model each of the start dates using a model with a semi-conjugate multivariate normal prior and multivariate normal likelihood. We choose 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 distribution of each of the start dates of dry and wet seasons are approximately normal as seen in Figure 3. Thus, we continue with the multivariate normal likelihood with confidence that it represents the data we have.

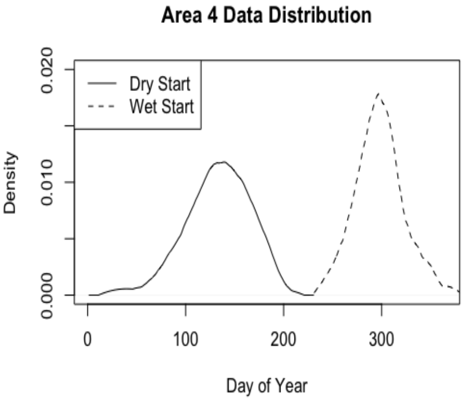

Figure 3: Kernel Density Estimate of the start and end of the dry season for Portland

We first define the semi-conjugate prior distributions as follows.

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

To determine the prior parameters, we let $\mu_0 = \begin{bmatrix} X_{min} \\ X_{max} \end{bmatrix} = \begin{bmatrix} 170 \\ 260 \end{bmatrix}$ and $\Lambda_0 = \begin{bmatrix} 30^2 & 0 \\ 0 & 30^2 \end{bmatrix}$ for all 9 area groups. The prior means are 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. To set the prior parameter for Σ, we choose the Σ_0 values by mimicking the methods used during our homework, where we let Σ_0 equal the *variance - covariance* matrices for each area group from the data. The parameters for Σ_0 of the nine regions are shown in Table 1.

In order to conduct Gibbs sampling, we need full conditionals of θ and Σ, which are shown below. By running Gibbs sampling on these full conditionals, we are effectively sampling from the posterior distribution

1	2	3	4	5	6	7	8	9
$\begin{bmatrix} 1154.93 & 179.97 \\ 179.97 & 618.95 \end{bmatrix}$	$\begin{bmatrix} 658.26 & 134.48 \\ 134.48 & 423.75 \end{bmatrix}$	$\begin{bmatrix} 1006.46 & 115.78 \\ 115.78 & 434.10 \end{bmatrix}$	$\begin{bmatrix} 1000.96 & 87.99 \\ 87.99 & 575.44 \end{bmatrix}$	$\begin{bmatrix} 993.37 & 236.99 \\ 236.99 & 405.11 \end{bmatrix}$	$\begin{bmatrix} 2943.59 & 630.29 \\ 630.29 & 728.15 \end{bmatrix}$	$\begin{bmatrix} 3574.88 & 884.72 \\ 884.72 & 874.46 \end{bmatrix}$	$\begin{bmatrix} 2221.71 & 210.85 \\ 210.85 & 500.14 \end{bmatrix}$	$\begin{bmatrix} 1974.74 & 250.49 \\ 250.49 & 684.75 \end{bmatrix}$

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

and so we can obtain posterior estimates of θ and Σ.

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

$$\begin{aligned} \mu_n(\Sigma) &= (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}) \\ \Lambda_n^{-1}(\Sigma) &= \Lambda_0^{-1} + n\Sigma^{-1} \\ \nu_n &= \nu_0 + n \\ \Sigma_n(\theta) &= \Sigma_0 + \sum_{i=1}^{n}(y_i - \theta)(y_i - \theta)^T \end{aligned}$$

3.2 Probability of Precipitation Model

We further develop our analysis by considering the probability of precipitation 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 and Discussion

4.1 Multivariate Normal Model

Using the semi-conjugate multivariate normal model described in section 3.1, we used a Gibbs sampler to generate samples from the posterior distributions of θ and Σ as well as the posterior predictive distribution $\begin{bmatrix} X_{min} \\ X_{max} \end{bmatrix} \sim MVN(\theta, \Sigma)$. When drawing samples from the posterior predictive distribution, we used the bound: $(X_{min}, X_{max}\epsilon(0, 365)$ to ensure that the sampled data points represents dates in a year. Because we used Gibbs sampler, we needed to take into account of the dependencies of the Markov Chains. This is important because issues may arise if the chain is stuck in one area of the distribution because of the previous sample. To account for these dependencies, we ran 20,000 simulations in order to obtain high effective sizes. On average, the effective sizes were: $\theta_{X_{min}} : 10,795, \theta_{X_{max}} : 7,229, \Sigma_{X_{min}} : 11,286, \Sigma_{X_{max}} : 8,906$. We then examined the autocorrelation plots and saw that it was fairly high, especially for the lag periods of 1 and 2 iterations. To resolve this issue, we thinned out the data by taking every tenth entry which resulted in 2,000 sampled values. Figure 4 shows the autocorrelation plots before and after thinning the data. We then analyzed the data using this thinned out dataset.

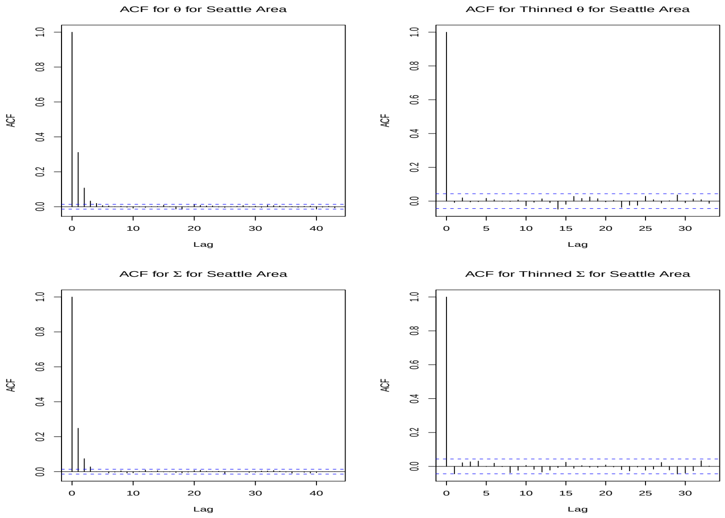

Figure 4: Autocorrelation plots for parameters before and after thinning data

We report summary statistics of θ, the posterior mean of the start date of the dry and west seasons, in Table 2. In particular we note the variability in the mean start dates of the dry season between the regions, which ranges from 149 - 162, and the increase in the mean start as we move towards the eastern regions. On the other hand, the end date of the dry season is much more consistent within and across regions, ranging from 180 to 189. This suggests larger weather patterns triggering the end of the dry season, which are consistent between years.

Region	Start Date Mean	Start Date Variance	End Date Mean	End Date Variance
1	152	24	280	17
2	149	20	288	18
3	153	22	284	15
4	154	23	282	22
5	153	17	289	12
6	158	30	283	27
7	158	28	284	28
8	162	25	289	20
9	160	26	284	30

Table 2: Posterior Mean and Variance of θ

Using the predictive posterior sample, we generated probabilities that an area has the earliest dry season and latest wet season in a given year. The probabilities are shown in Table 3. Looking at the probabilities, we see that the Columbia Plateau and Selkirk Mountains are most likely to have the earliest dry season. This aligns with the climate of the area since the Columbia Plateau is somewhat arid because of the rain shadow of the Cascade Mountains (but selkirk is wet?). However, it is interesting because both areas are quite far from each other (approximately 400 km), which suggests that the precipitation patterns do vary because surrounding areas are not as likely to have the earliest dry season. Looking at the lengths of the dry season in Table 3, there is quite some variability, with a difference of 20 days of the longest and shortest periods The Oregon Coastline and Cascade Range have the longest dry season while the Eastern Oregon area has the shortest dry season.

Also of interest is the beginning of dry, summer weather across the Pacific Northwest. In addition, a very likely beginning of the dry season is later in the generally drier, Eastern part of the region. This may be due to the drier nature of that region, so distinctly different weather patterns may not develop until later in the summer.

Region	Earliest Dry Season	Latest Wet Season	Length of Dry Season
1	0.099	0.074	129
2	0.074	0.032	138
3	0.091	0.057	122
4	0.087	0.081	127
5	0.073	0.068	136
6	0.180	0.198	123
7	0.179	0.200	121
8	0.109	0.145	129
9	0.110	0.147	119

Table 3: Area with probabilities of earliest and latest start times as well as length of dry season

5 Conclusion