Final Report: Rainfall Trends in Mumbai, India

Heleyna Tucker (heleyna)

December 17, 2024

Introduction

Finding ways to model rainfall over time is important to summarize the magnitude of change and whether or not there are obvious trends as months or years go on. Highlighting these trends can bring up important discussion of climate change and possible solutions. In this paper, Bayesian linear regression and hierarchical modeling will be used to model the annual and monthly rainfall trends for Mumbai, India and quantify these trends. Mumbai has annual monsoon seasons, which brings in the method of separating monsoon (June-September) and non-monsoon months and analyzing their trends separately. Increasing rainfall over time could be caused by a number of factors. Finding evidence of these trends and their magnitude is necessary to start important conversations as to how these trends are caused.

Dataset

For this project, I will be looking at Mumbai monsoon rainfall (in mm) data from 1901-2021. This data was collected monthly. To get a good understanding of the dataset, visualizations and summaries were done. Some months, specifically the non-monsoon months, did contain zero values for rainfall. Additionally, overall variability in rainfall per month was looked at, showing that the monsoon months had the most variability, having a standard deviation ranging from 200-300 mm of rainfall. There were also some extreme and minimal rain years, shown in Figure 8 (in appendix), with the past 20 years notably having 3 rainfall years that were considered extreme. This is important to consider, especially for linear regression. High variability could lead to high residuals (how far the predicted value is from the observed). Figure 1 showcases the dominance of monsoon rainfall in this dataset. Figure 2 allows a deeper understanding of both the monsoon months and non-monsoon months. Here, we can see monsoon months have around 500 mm of rain, on average with some months reaching 1000 mm. Non-monsoon months range from 0 - 500 mm.

To gauge the overall trend of monsoon rainfall over the years, some preliminary graphs and a simple linear regression model were done. Figure 3 showcases the variation of the data while also capturing the linear trend over the years. These initial data findings pose the question of whether there has been an increase in rainfall over the years and how it can be modeled. Additionally, some hierarchical techniques could be used to separate the trend of monsoon and non-monsoon months.

Methods

The first area of interest was to assess whether total rainfall during the monsoon months has increased over the years. To investigate this, Bayesian linear regression was implemented using the rstan package (Stan Development Team). The model included two key parameters: α (intercept) and β (slope). The likelihood of the model was assumed to be Gaussian, reflecting the observed distribution of total rainfall.

The data from the monsoon months showed a mean total rainfall of approximately 2000 mm and a standard deviation of 300 mm per year. Consequently, the likelihood function for total rainfall per year was defined as $Normal(\alpha + \beta * Year, \sigma)$. where α is the intercept, β is the slope, and σ represents the standard deviation of the residuals. To enhance model convergence, the year variable was normalized by subtracting the minimum year (1901) from all observations, resulting in a normalized range starting at 0. The following priors were chosen based on the data provided: α was normally distributed with parameters 2000, 300 and β was assumed to have a slightly positive trend, having a distribution of Normal(5, 10) and σ was chosen to be Cauchy(0,5). Figure 4 presents the posterior predictive checks that emphasize the components were correctly implemented for the model.

To assess the predictive capability of the model, a train/test split was performed, and performance metrics such as root mean square error (RMSE) were calculated to evaluate the model's ability to predict future rainfall trends.

The second area of interest focused on capturing seasonal patterns across the entire dataset. The data was restructured into a format ('rainfall_long2') where monthly rainfall values were recorded alongside a time variable representing the number of months since January 1901. For example, January 1901 was encoded as 1, and January 1902 was encoded as 13. Additionally, a categorical variable, 'season', was introduced to indicate whether each observation corresponded to a monsoon or non-monsoon month. A hierarchical Bayesian model was made with rainfall per month as our response variable and Time as our predictor. The goal of this model was to separate the seasonal trends while modeling for the whole dataset. The likelihood for this model was Normally distributed with parameters $(\alpha[Season[n]] + \beta[Season[n]] * Time[n], \sigma)$. Other likelihoods and different set-ups (non-linear) could be explored, explained in the results section, however a normal likelihood with linear parameters was chosen for better understanding. Summary statistics were used on the monsoon(1) and non-monsoon(2) months to obtain the normal priors of α_1 with parameters (500, 300) and α_2 with parameters (14, 50). The beta priors were set to normal (5,10) assuming a slightly positive trend for both seasons. Sigma was again chosen to be Cauchy(0,5). It is important to note that non-monsoon data points seem to have a lot of 0 mm measurements (564/968) which is most likely the reason for the high variability and could affect the accuracy of our model and beta estimate.

Posterior predictive checks (Figure 5) indicated reasonable fit for the model but highlighted some discrepancies. For example, the predicted rainfall density did not match the observed peak, suggesting areas for future model refinement. Possible extensions include exploring non-linear trends or alternative likelihood formulations to better capture variability in the data.

Results

The first linear regression model looked into the main question of whether rainfall shows an obvious increase or decrease throughout the years, specifically within the monsoon seasons in Mumbai, India. Figure 6 showcases the trends we see for this data. The main area of interest is that beta is a positive value, with its confidence interval spanning only positive values. This gives researchers evidence that over the years, the rainfall has increased. The slope estimate is 4.39, which means that with every one-year increase in year, the rain measurement goes up by 4.39 mm, on average. The value for sigma (measuring residual variability) is around 500. This suggests that our

predicted and observed values are quite far from each other. In addition, the predictive power of this model was tested. Modeled by a subset of training data (all years at or below 1990) and test data (all years above 1990). This model showed similar results in Figure 9 (in appendix) and had an RMSE of around 500. This high positive value indicates that our model is not performing very well. A quadratic term was also explored with little to no notable results. The high RMSE could be due to the incorrect choice of likelihood or modeling technique. For example, a linear trend does not capture variability of data well. If I were interested in predicting future rainfall, more predictors would be needed than just yearly trends alone.

The hierarchical model results includes all months and splits them into monsoon and non-monsoon, exploring the monthly trends throughout the years 1901 - 2021. The results for this model are shown in Figure 7. Key features to highlight in these results are that the non-monsoon months seem to be quite difficult to evaluate. With a small α_2 (intercept) and a basically 0 β_2 (slope) value, there seems to be little change over time. Possible ways to investigate this further are explained in the next section. Additionally, we see the same trend as previously in this hierarchical model, that monsoon months trend upwards on rainfall measurements. This is shown by the positive β_1 (slope) value and the exclusively positive confidence interval. Although small, the estimate of 0.09 for the slope is based on a monthly scale and, according to our confidence interval, is most definitely trending upwards per month. Overall, these results give researchers a model that enhances the understanding of rainfall patterns over the years and months of monsoon months specifically. Additionally, the sigma values seem to be quite high (about 175) bringing up a possibility to expand or do more research on how these models could better detect this variability.

Limitations/Conclusion/Future Work

Overall, this project helped me learn more about the Bayesian framework and the importance of choosing good priors. Initially, I had the default priors for these models but learned that including prior knowledge was worth it. From these results, it can be inferred that monsoon rainfall is increasing over time in Mumbai, with a positive beta coefficient each year, the trend seems to be increasing steadily. This could be a possible area of concern for the general public and researchers. If I had more time, I would find another way to predict rainfall with another dataset. For example, temperature and wind measurements would add a lot of predictive power, more than only measuring trends over the months/years. Additionally, I did find an article that discussed the implementation of Fourier parameters in their model (Lall, Lima). They had similar data and looked to have had better results that fit the varied nature of the observations. A linear model with normal likelihood and priors seem to have not captured the variability in the data, therefore this new approach could lead to more precise modeling.

Figures

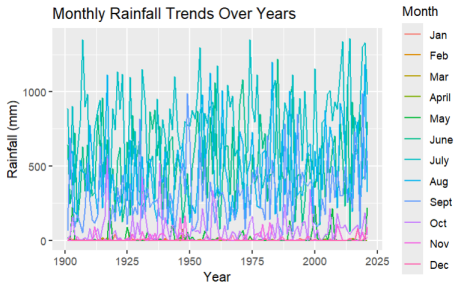


Figure 1: Line Plot of rainfall patterns over the years colored by month.

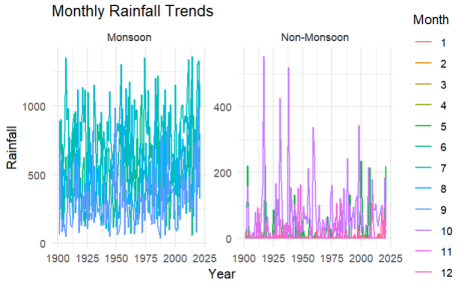


Figure 2: Line Plot of rainfall patterns over the years colored by month, separated by whether the month is considered a monsoon month (June-September).

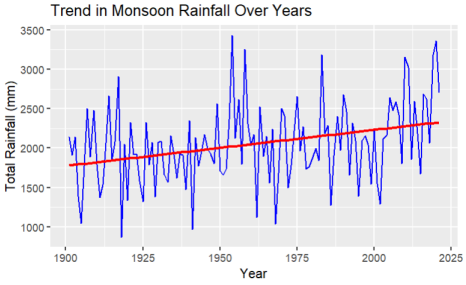
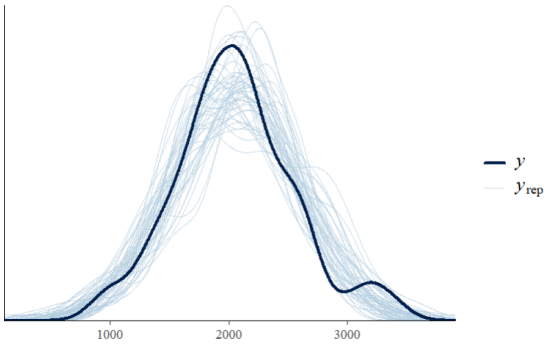
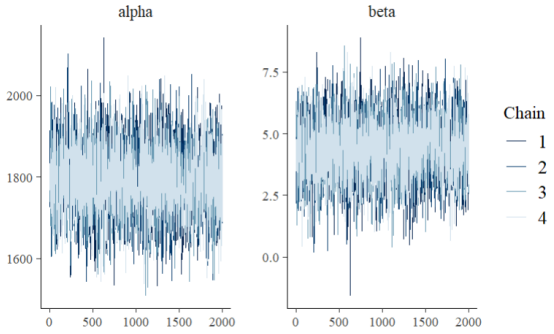


Figure 3: Line Plot of rainfall patterns over the years for monsoon months with overlaid linear model with year as the explanatory and total rainfall for monsoon months as the response variable.

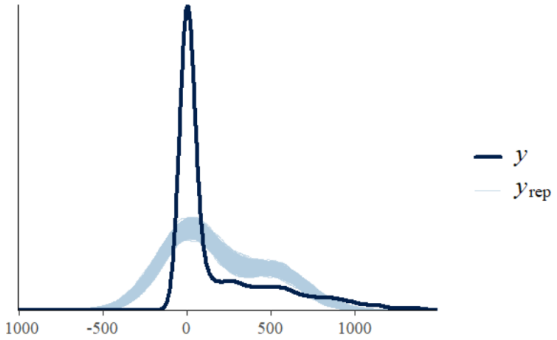


(a) Posterior predictive density plot.

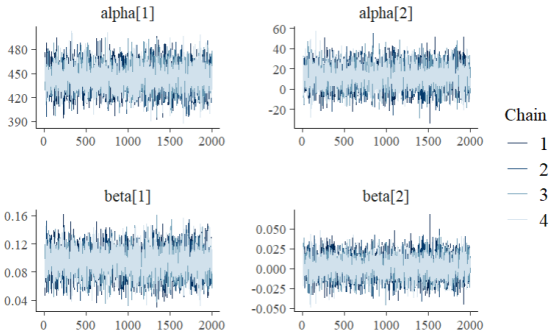


(b) Trace plots for parameters.

Figure 4: Posterior predictive checks for the Bayesian linear regression model.



(a) Posterior predictive density plot.



(b) Trace plots for parameters.

Figure 5: Posterior predictive checks for the Bayesian hierarchical linear regression model.

```
Inference for Stan model: anon_model.
4 chains, each with iter=4000; warmup=2000; thin=1;
post-warmup draws per chain=2000, total post-warmup draws=8000.

        mean se_mean    sd    2.5%     50%   97.5% n_eff Rhat
alpha 1793.43    1.42 81.48 1635.98 1792.86 1957.51  3275    1
beta     4.39    0.02  1.19    2.02    4.41    6.69  3230    1
sigma  480.68    0.45 30.68  425.21  479.72  546.57  4612    1

Samples were drawn using NUTS(diag_e) at Sat Dec 14 13:47:05 2024.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Figure 6: Bayesian Linear Regression results for predicting total rainfall with year.

```
Inference for Stan model: anon_model.
4 chains, each with iter=4000; warmup=2000; thin=1;
post-warmup draws per chain=2000, total post-warmup draws=8000.

           mean se_mean    sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
alpha[1] 444.71    0.22 15.81 413.75 434.19 444.60 455.25 475.49  5161    1
alpha[2]  12.56    0.15 11.13  -9.11   4.96  12.59  20.06  34.14  5617    1
beta[1]    0.09    0.00  0.02   0.06   0.08   0.09   0.11   0.13  5003    1
beta[2]    0.00    0.00  0.01  -0.02  -0.01   0.00   0.01   0.03  5794    1
sigma    173.44    0.04  3.29 167.21 171.18 173.37 175.62 180.16  6329    1

Samples were drawn using NUTS(diag_e) at Sat Dec 14 15:12:25 2024.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Figure 7: Bayesian Hierarchical Linear Regression results for predicting total rainfall with year separated by monsoon season.

Appendix

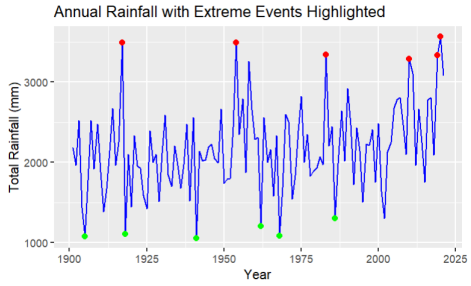


Figure 8: Line Plot of rainfall patterns over the years for all months with extreme points colored.

```
Inference for Stan model: anon_model.
4 chains, each with iter=4000; warmup=2000; thin=1;
post-warmup draws per chain=2000, total post-warmup draws=8000.

        mean se_mean     sd    2.5%     50%   97.5% n_eff Rhat
alpha 1719.80    1.69 92.30 1537.13 1721.09 1898.53  2974    1
beta     5.25    0.03  1.81    1.75    5.22    8.85  2940    1
sigma  474.26    0.54 35.91  410.03  471.69  550.91  4403    1

Samples were drawn using NUTS(diag_e) at Fri Dec 13 20:38:31 2024.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Figure 9: Bayesian Linear Regression prediction results for predicting total rainfall with year.

Citations

Stan Development Team. 2018. Stan Modeling Language Users Guide and Reference Manual, 2.32.6. https://mc-stan.org

U. Lall and C. H. R. Lima, Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA. (ula2@ columbia.edu; chr2107@columbia.edu)

Dataset: https://www.kaggle.com/datasets/macaronimutton/mumbai-rainfall-data