An Analysis of Air Quality Changes and Related Chronic Disease Rates in the United States, 2000-2019

Anthony Carlos Mendoza
University of California, Berkeley

CJ Jin
University of California. Berkeley

Kenneth Choi
University of California, Berkeley

Selena Zhao
University of California, Berkeley

Abstract

Recent novel research into particulate matter presence in the atmosphere shows that these particles present a significant danger to the human respiratory system. Particulate matter (PM2.5) is described as particles that possess an aerodynamic diameter of less than $2.5\mu m$. These particles get absorbed into the human body through the respiratory system where they stimulate pulmonary inflammation, which consequently leads to numerous respiratory-related illnesses. PM2.5 is attributed to the death of nearly 4 million people globally by these illnesses, namely, asthma, chronic lung disease, cancers, cardiopulmonary diseases, and other infections. In this work, our team analyzes the changes in particulate matter presence in the atmosphere through observations in the United States, from 2001-2019. We show that the trends are positive, but deferentially so, with certain regions fairing better than others. Next, we utilize machine learning methods to predict future outcomes of the levels PM2.5 to inform the reader on areas which require more resources to better minimize the risk of PM2.5 negatively impacting people of those areas. Next, we build a Bayesian Hierarchical model to understand the chronic disease indicators, namely asthma, across the regions being studied.

1 Research Question: can we fit a parametric and/or non-parametric model to best capture the nuances of the changing PM2.5 levels and accurately make predictions for future outcomes?

1.1 Introduction

Over the past two decades, the levels of particulate matter (PM2.5) have fluctuated unevenly in the United States. To understand these changes, we utilize data provided by the CDC Air Quality Indicator (AQI). We opt to aggregate the states in the U.S. by their census-designated regions (West, South, Northeast, Midwest) due to computational constraints. By aggregating the states into their specific regions, we see from figure 1 and figure 2 that there is not much loss in generality, as each states tends to behave similar to the other states in the region, citing similar levels of PM2.5 in the beginning year, 2000, the end year, 2019, and the general seasonal and periodic trends, which we rigorously examine later in this work. The notable exception to this caveat is the state of California, which perhaps most uniquely presents as an outlier to the other Western states in its region, encompassing levels of PM2.5 which are notably higher than the other Western states, but arrives to a similar outcome level in 2019 after an impressive decrease in PM2.5 through the nearly two decades observed. This behavior is captured in the aggregation, highlighted by the width of the line corresponding to the Western states in figure 2.

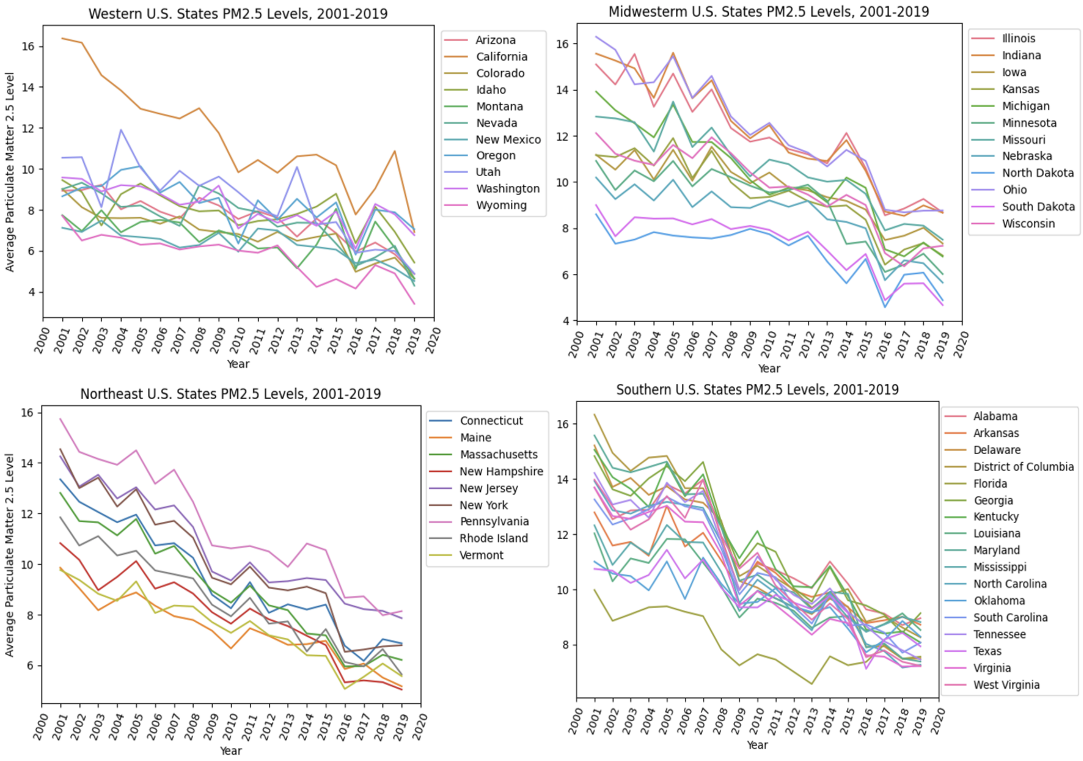


Figure 1: State changes in PM2.5 levels, 2001-2019

Our team opted for an approach that models outcomes rather than inputs for a simple reason: given the myriad amount of confounding variables likely persistent to explain the PM2.5 data (i.e. GPD of each state, GDP of each region, regional and state-specific environmental protection policies, industry presence, composition of state and regional economies from manufacturing/services basis, etc.) a macroeconomic approach of utilizing causal inference seemed infeasible and any result presented would be dubious for the nature of the complexity involved in attempting to diagnose, explain, and predict economic-based outcomes, cannot yet be captured in a single paper without access to copious amount of related data. Thus, to abstract the economic-based complexity, we opted for an approach presented in recent machine learning research on this topic: a time-based modeling approach that predicts future outcomes from current and past outcomes. This approach reliably produces a result that is not encumbered by confounding factors: it assumes that the present and past factors may or may not be present in future outcomes, its prediction power is independent of them in its forecasts. We begin with a time-series based parametric modeling: fitting ARIMA and SARIMA models, evaluating prediction power, then move to GLMs Linear Regression model as the baseline, then Lasso (L1 Regularization) and Ridge (L2 Regularization), and finally end with the non-parametric LSTM neural network model. In all models, we compute time-based lags and treat them as additional predictors into our models.

1.2 Exploratory Data Analysis (EDA)

As mentioned above, our team retrieved the PM2.5 data from the CDC, in four data sets each containing five years of data for the PM2.5 levels observed for the specific time period. The granularity of the data, before grouping, was annual PM2.5 level for a state. As we grouped the PM2.5 annual level for each state by its region, first by creating a function to identify the respective associated region for the state, then by aggregating the values by the mean, such that each row in the new data frame contained four columns, the granularity inevitably changed to each row representing the year and the subsequent columns representing the PM2.5 level for that region. As our data was already well-processed, with no missing values, we proceeded to analyze the structure of the outcomes.

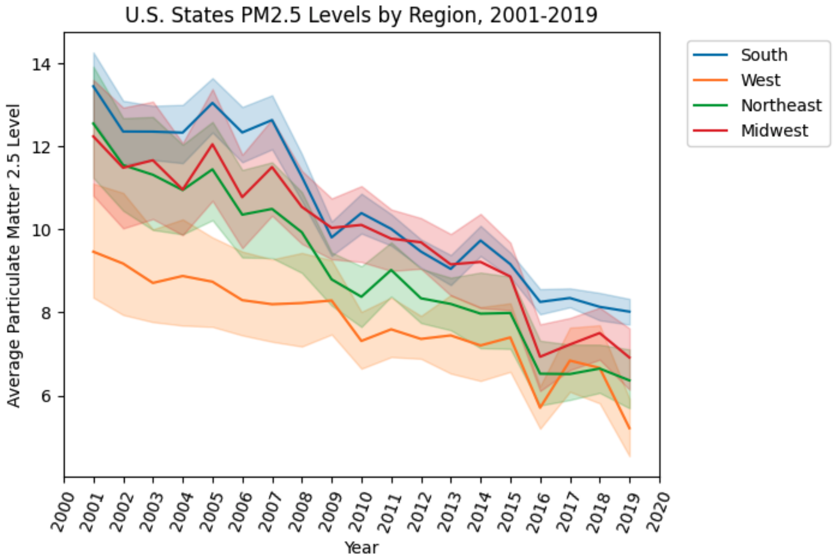


Figure 2: Regional changes in PM2.5 levels, 2001-2019

1.2.1 West Region Data Analysis

We begin this parametric analysis of data in the West region of the United States by trying to identify the structure of our time-based data. In figure 3, we present the findings of seasonal decomposition. The left images present the decomposition itself, where we visually determine the following: the west region has a negative trend, something we suspect will be common for all four regions of the United States. We see there is no present seasonality in the data, allowing us to surmise that the any subset of time periods do not reflect common fluctuations. From the residual plot, the lack of noise is notable, a testament to pre-processing done beforehand. The right two plots indicate the partial autocorrelation function (PACF) and the autocorrelation function (ACF) of the data. First, the PACF provides insight into the correlation between the observations at time, t, and the observation at a respective lag, $t + h$, allowing us to visually inspect the order of the autoregressive (AR) component to later utilize in constructing our model, as well as providing us a guide into the optimal amount of lags to construct for our GLM and LSTM model . Evaluation of the ACF provides us the moving average (MA) component for our model. To understand the stationarity of our model, we utilize the ADFuller Test and see the results (figure 4) to determine our data is likely not stationary, due to the high p-value, $p - value = 0.99$, which exceeds our threshold of $\alpha = 0.05$.

1.2.2 South Region Data Analysis

Next, we move to parametric analysis of data in the South region. In figure 5, we present the findings of seasonal decomposition. From these, we notice similarities to the West region, namely, there is a negative trend, albeit with a different structure and no seasonal component observed through the time period. Also, we similarly note the lack of noise as attained from the residual plot. The PACF and ACF plots show similarities to the West region, but the similarities stop there: from the ADFuller test, we see that the South region data is likely stationary, with $p - value = 0.007$, not exceeding our threshold of $\alpha = 0.05$, alluding to a possibility of fitting a GLM to outperform an ARIMA or SARIMA model, something we examine below in our modeling section.

1.2.3 Northeast Region Data Analysis

Third, we move to parametric analysis of data in the Northeast region. In figure 6, we note the following: a negative trend as has been observed with both the South region data and West region data mentioned above. Again, we note that there is likely no seasonality present in the data and little to no noise observed. The PACF and ACF allude to a potential similar modeling hyperparameters to the South Region, which is supported by the likely stationarity present in the data, which we observe

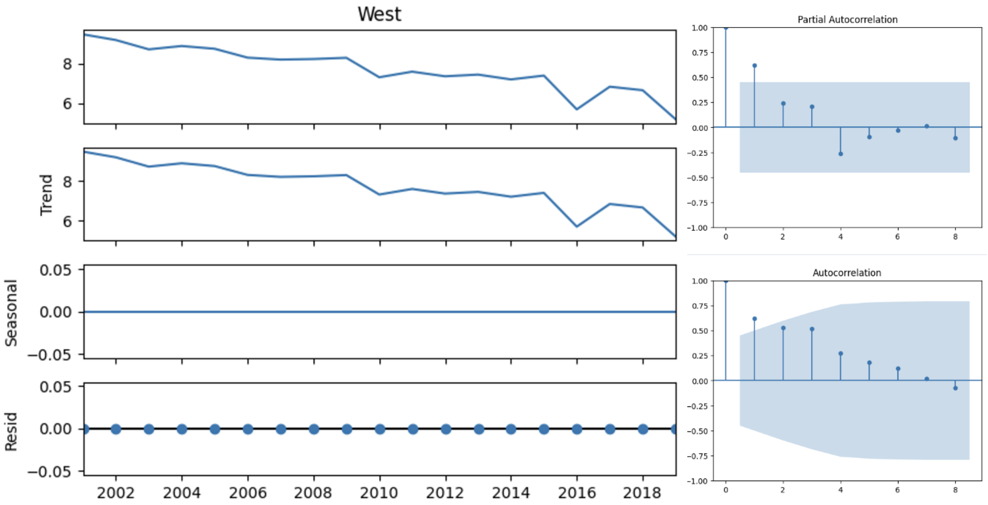
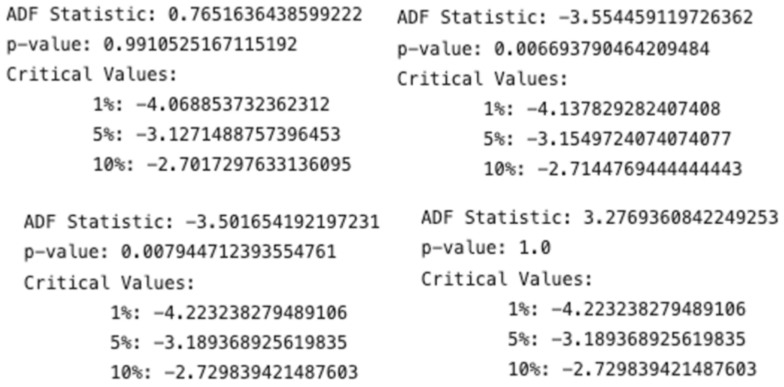


Figure 3: West Region Data Analysis



```
ADF Statistic: 0.7651636438599222
p-value: 0.9910525167115192
Critical Values:
        1%: -4.068853732362312
        5%: -3.1271488757396453
        10%: -2.7017297633136095
```

```
ADF Statistic: -3.554459119726362
p-value: 0.006693790464209484
Critical Values:
        1%: -4.137829282407408
        5%: -3.1549724074074077
        10%: -2.7144769444444443
```

```
ADF Statistic: -3.501654192197231
p-value: 0.007944712393554761
Critical Values:
        1%: -4.223238279489106
        5%: -3.189368925619835
        10%: -2.729839421487603
```

```
ADF Statistic: 3.2769360842249253
p-value: 1.0
Critical Values:
        1%: -4.223238279489106
        5%: -3.189368925619835
        10%: -2.729839421487603
```

Figure 4: ADF Statistics, top row: West, South, bottom row: Northeast, Midwest

through the ADFuller test, which produces a $p-value = 0.007$ which does not exceed our threshold of $\alpha = 0.05$.

1.2.4 Midwest Region Data Analysis

Lastly, we examine the Midwest Region data in figure 7, and note the similarity in structure of the downward trend that tends in similarity to the West Region moreseo than either the South Region data or the Northeast Region data. The ACF also points to a potential higher order for the moving average (MA) component, something we also observed for the West Region, whereas the South Region data and Northeast Region data likely have lower MA orders. The PACF of all four regions seems to be shared, though we conduct hyperparameter tuning in our modeling section below to examine this more closely. Finally, much like the West Region data, we note the ADFuller Test provides a $p-value = 1.0$, which points to non-stationarity in the data, something only observed with the West Region data thus far.

1.3 Methods

1.3.1 Parametric Modeling: ARIMA

Our goal in this section to predict the future levels of PM2.5 for each region provided previous annual outcomes spanning 2001-2019. Provided the time-based nature of our data, we make use of

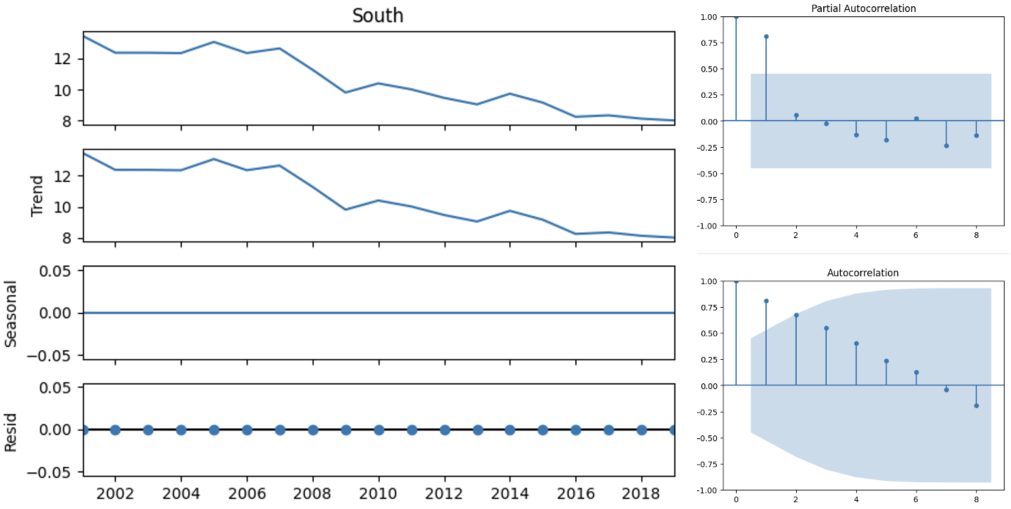


Figure 5: South Region Data Analysis

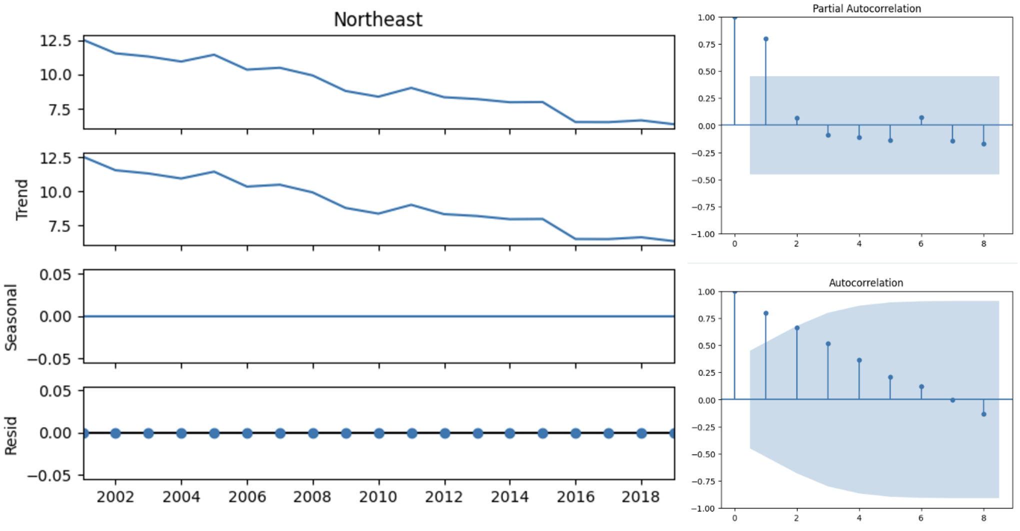


Figure 6: Northeast Region Data Analysis

time-dependent methodologies and first opt to construct a traditional timeseries-based parametric model, ARIMA. We create a modeling pipeline that does the following: accepts each regional data, performs time-based grid search hyperparameter tuning to find the optimal parameters (the results of which are supported by our findings through our PACF and ACF plots above), fit the optimal models, and evaluate using RMSE. To improve predictive power, we also construct exogenous features that represent the lag at $t + 1$ and $t + 2$, respectively. This approach is common in timeseries modeling for lagging the series makes the past outcomes contemporaneous with the predictions. Inherently, this induces some assumptions, namely, that there may be stationarity in our data, which we observed above through the ADFuller test restuls for two of our four regions. We also assume that there is independence in the residuals, which we show later in our results to hold. Lastly, we assume there is a linear relationship between our variables. Of note, though we create a SARIMA model for each region (4x additional models), given our findings of the lack of seasonality in our data for all regions, the ARIMA models outperform each of their SARIMA counterparts so we do not include them here for brevity's sake.

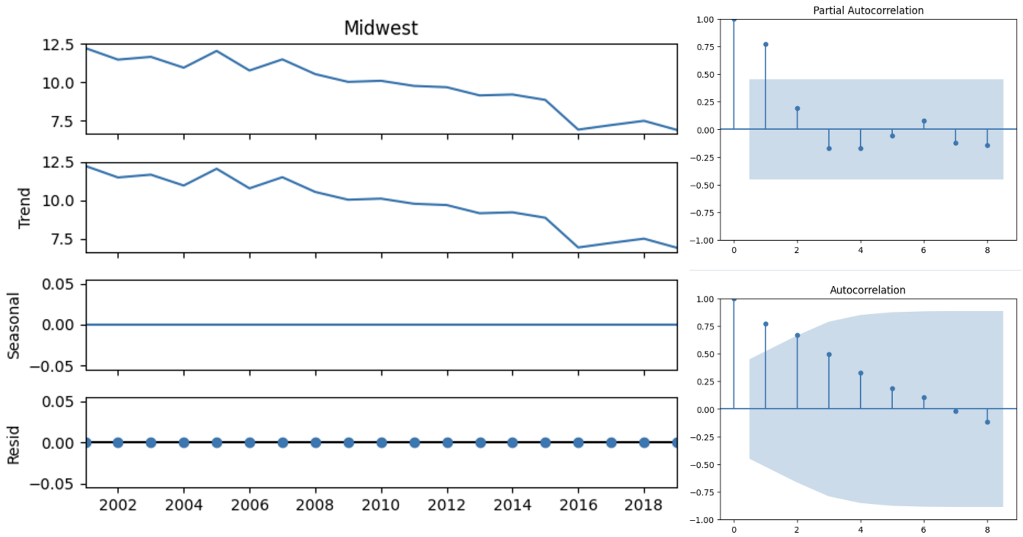


Figure 7: Midwest Region Data Analysis

1.3.2 Parametric Modeling: Generalized Linear Models (GLMs)

Provided our findings in our exploratory data analysis, where we found that the West Region data and the Midwest Region data presented non-stationarity, the use of an ARIMA model to capture the non-stationarity proved fruitful, with the appropriate use of p, d, q parameters to capture the moving average, q, autoregressive component, p, and differencing factor, d. In analyzing our other two regions, namely, the South Region and Northeast Region, we observed stationarity in these datasets, thus we hypothesize that a GLM with regularization would likely outperform our timeseries-based models. To approach this modeling problem, we assume the following: our data are linear, which we accept as we did above. Further, we assume our results of stationarity holds for our South and Northeast regions, and our our results of non-stationarity holds for our West and Midwest regions. As with our parametric models above, we also assume independence in residuals. While this assumption is the most difficult to accept given that temporal dependence is intrinsic to time-based data, we accept the assumption as we similarly create lags as additional features and utilize the appropriate number depending on the ACF plots for each dataset as discussed in our EDA section.

Similar to our ARIMA model above, we construct a pipeline that does the following: first, it adds two additional features to lag our data. Next, we use Sklearn's TimeSeriesSplit model given the inherent temporal structure of our data which requires a careful consideration of chronological order. This method uniquely facilitates the meticulous separation of training and test sets, ensuring that future observations do not leak into the past during model evaluation. The nuanced temporal stratification achieved by TimeSeriesSplit, recognizing the inherent dependency structure within sequential data, results in a more apt evaluation framework for time series forecasting models. This tailored approach minimizes the risk of information leakage, thereby enhancing the robustness and reliability of predictive performance assessments in time series analysis. Traditional K-fold CV is not appropriate given, again, the temporal nature of the data. A visualization is provided in figure 8 (from Rob J Hyndman and George Athanasopoulos). For our case, as we utilize 2 lags, therefore our predictions are made for 2 time steps ahead.

1.3.3 Nonparametric Modeling: LSTM Neural Network

Inherent in any machine learning prediction problem that abstracts confounding factors by focusing on temporal outcomes is a level of complexity that might not be explained solely through the use of parametric techniques that solely seek to model the outcomes based on some parametric assumptions. As such, we construct a LSTM network. Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks (RNNs), have recently emerged in literature as a preeminent choice for time series prediction tasks. This preference likely stemmed from their adeptness in grasping intricate temporal dependencies inherent in sequential data, offering a formidable solution to challenges

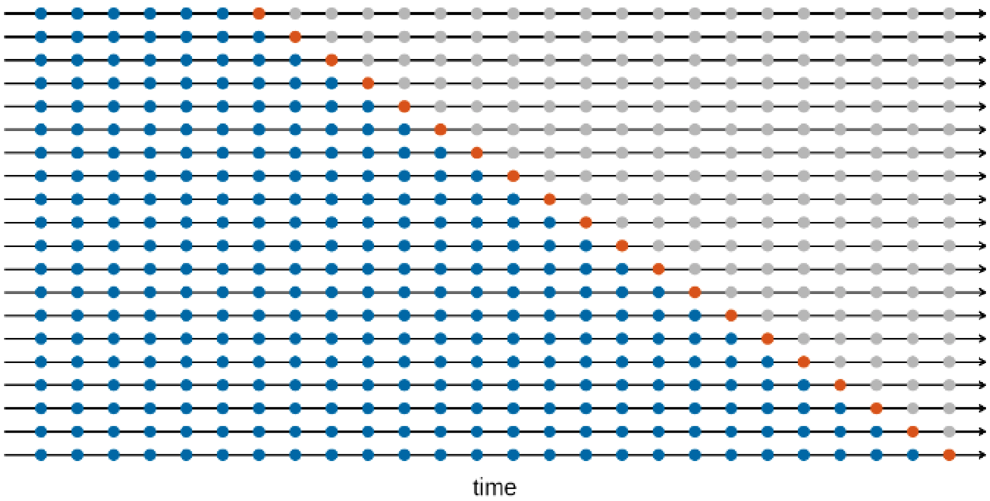
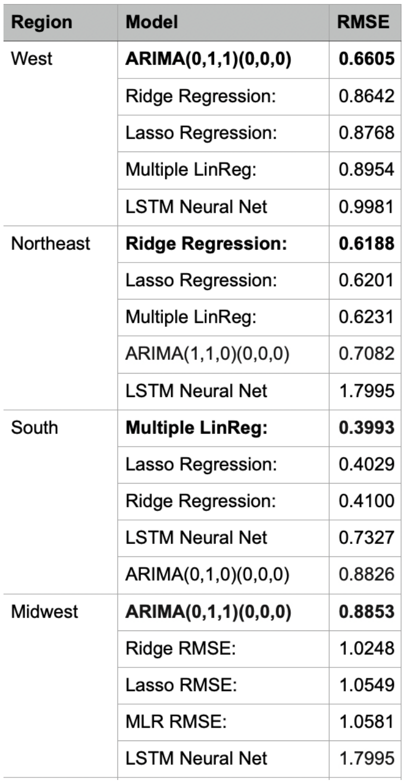


Figure 8: HA: Time Series CV Visualization

encountered by traditional models. Through the utilization of memory cells, which are specialized units designed to retain information over extended sequences, LSTMs have proven to be able to discern and store essential temporal patterns, empowering the model to recognize and exploit long-term dependencies within time series data. However, as we quickly learned, their predictive power is also closely coupled with large computational requirements, which unfortunately we were unable to sustain on our local machines. Though this proved to not be a problem in the base construction of 4x LSTMs (one for each region), this proved to be a problem in our optimization efforts, thus we had to resort to constructing our pipeline to build the LSTMs using the same un-tuned parameters for each of the models. Our assumptions for using the LSTM were similar to some of those made above, albeit fewer were made (due to the nonparametric nature). Namely, we assumed temporal dependence, as LSTMs are designed to capture temporal dependencies in sequential data, which we accept easily since this is a given with time-based data. Next, we assume some temporal stability, since, while LSTMs are able to handle stationarity, they do not perform well with seasonal or trend components. We accept this assumption for while stationarity was observed in two of our four regions, and a negative trend was observed in all four regions, these were still able to be captured by our parametric techniques above.

1.3.4 Results and Discussion

In general, our results allude to a point we predicted earlier: for the regions where the data was non-stationary, the ARIMA model outperformed the GLM models. We predicted this earlier as we figured that the ARIMA model would be able to (through the differencing component) capture the non-stationarity in its fitted values and thereby produce better predictions. For the West Region, the ARIMA(0,1,1) model performed better on our test set than the GLM and LSTM models. This is also true for the Midwest region. In the prediction plots in figure 10, we see the extent of the differences: the West and Midwest models for the GLMs are not able to capture the structure of the test set, which is not true for the South and Northeast regions, which show the GLMs were able to capture the structure of the test well. To this, we note that for these two regions, the GLMs outperformed the ARIMA model: with the South having the base multiple linear regression model barely improve upon the RMSE than the Lasso regression model (0.3993 to 0.4029, respectively), and the Northeast region having the Ridge regression model outperform the other models. For all models, we opted to utilize the root mean squared error (RMSE) to test performance. Our nonparametric LSTM underperformed the parametric models in all regions. However, as alluded to in our methods section above, this is more likely due to the computational limitations of our machines moreso than the model itself. As we were unable to perform hyperparameter tuning fot the LSTM, we were limited to fitting only a base model for all regions, a testament to relatively weak performance compared to our parametric models.



Region	Model	RMSE
West	**ARIMA(0,1,1)(0,0,0)**	**0.6605**
	Ridge Regression:	0.8642
	Lasso Regression:	0.8768
	Multiple LinReg:	0.8954
	LSTM Neural Net	0.9981
Northeast	**Ridge Regression:**	**0.6188**
	Lasso Regression:	0.6201
	Multiple LinReg:	0.6231
	ARIMA(1,1,0)(0,0,0)	0.7082
	LSTM Neural Net	1.7995
South	**Multiple LinReg:**	**0.3993**
	Lasso Regression:	0.4029
	Ridge Regression:	0.4100
	LSTM Neural Net	0.7327
	ARIMA(0,1,0)(0,0,0)	0.8826
Midwest	**ARIMA(0,1,1)(0,0,0)**	**0.8853**
	Ridge RMSE:	1.0248
	Lasso RMSE:	1.0549
	MLR RMSE:	1.0581
	LSTM Neural Net	1.7995

Figure 9: Modeling Results

While our results allude to optimistic predictions, more careful analysis is required to determine several things. First, examining the relationship between the stationarity of the data for each region and the predictive power of the GLMs vs the ARIMA model is required to understand the true relationship, if any, between deciding on which model to use for future observations when data presents stationarity or non-stationarity. Next, the LSTM needs to be trained on a machine with more computational power to truly encapsulate the potential of the model. In our uncertainty, we conduct the standard error estimate:

$$\hat{\sigma} = \sqrt{\frac{1}{T-K-M}\sum_{t}^{T} e_t^2}$$

and use this to build a 95% confidence interval for our GLM models,

$$\hat{y}_{t+h|t} \pm 1.95\hat{\sigma}$$

and plot this in figure 12. We see our predictions are within the 95% C.I. for most of the 12 models, showing our predictions are being made with high confidence. However, some models perform better than others. The MLR model is able to capture the true values with a high confidence in all regions. The Ridge model performs better for the models with stationarity, as with the Lasso model. In general, the frequentist approaches we utilized allowed us to more easily construct our models and quantify their uncertainty, therefore we would likely opt and recommend using these (and only distinguishing between ARIMA and GLMs after stationarity tests) over the nonparametric techniques, due mostly in part to computational requirements and interpretability of the models. Our frequentist approach provides us with depth in analyzing why models perform better than other given the structure of the data, as we discussed above. For our nonparametric models, the lack of interpretability and additional complexity prove these models difficult to utilize within this context. Given the time-based component inherent in our data, a Bayesian approach would not have been as computationally feasible, and likely not as appropriate, as we noted above with our nonparametric LSTM model. Further, placing a prior belief on the distribution would have introduced a heightened level of uncertainty which likely would not be appropriate given the time-dependence involved in this problem. In general, the frequentist parametric models outperform nonparametric techniques since the problem involved is a time-dependent parametric problem, and only compounds so with the introduction of lags as predictive features.

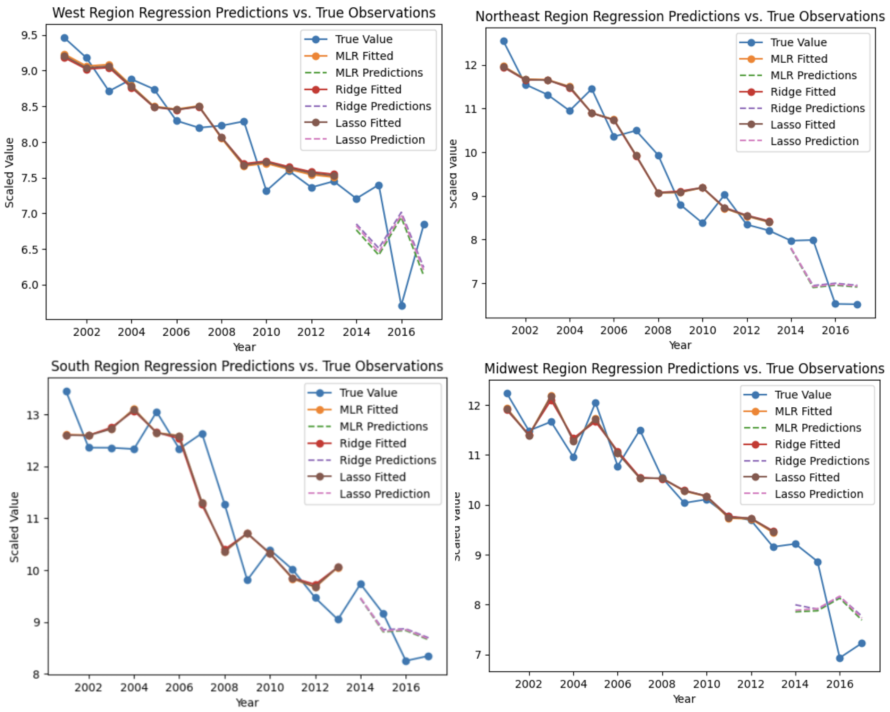


Figure 10: Parametric Modeling Plots

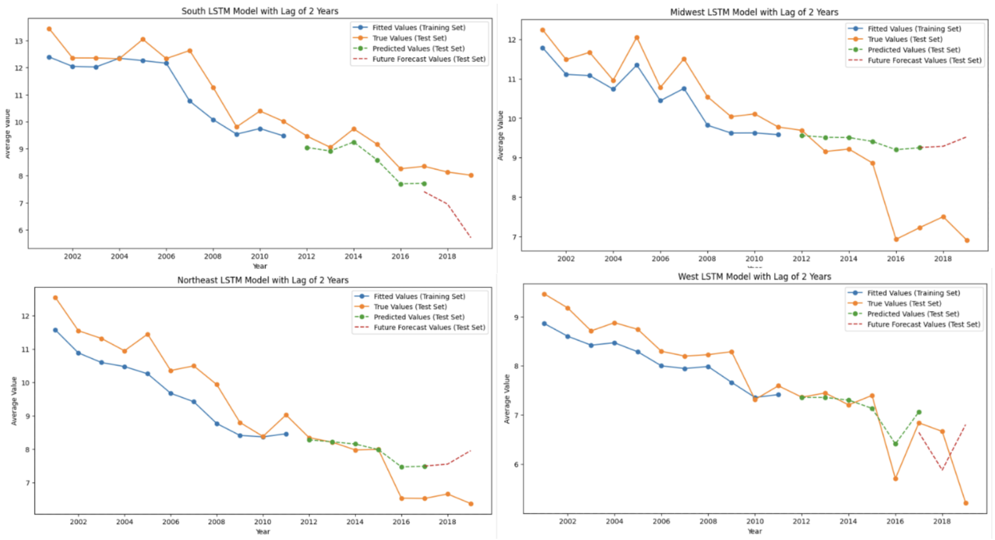


Figure 11: Nonparametric Modeling Plots

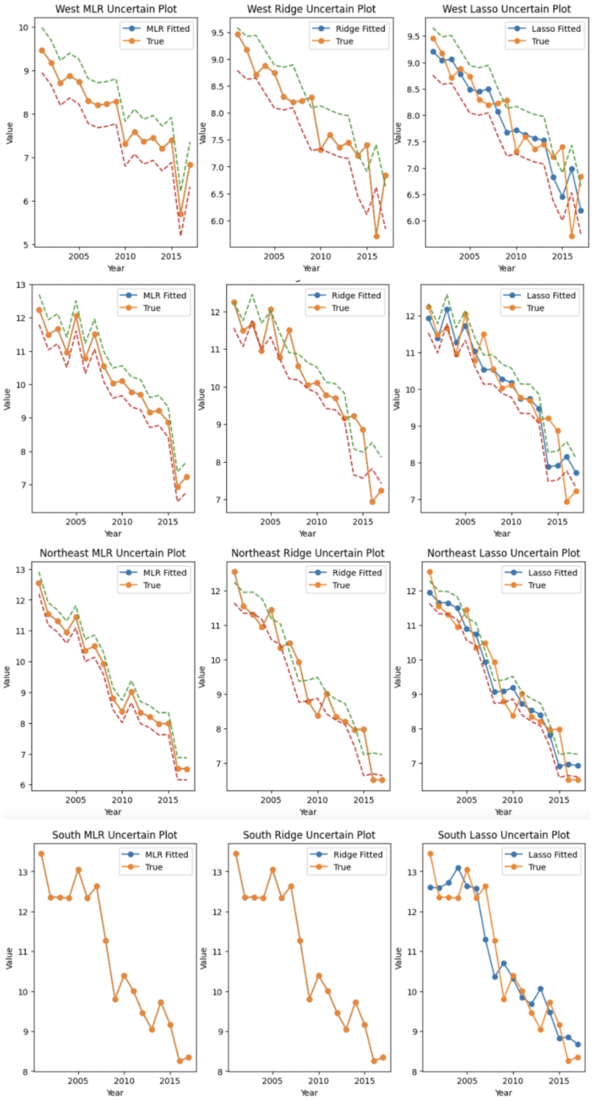


Figure 12: Modeling Uncertainty Plots

2 Research Question: can we examine the relationship between asthma rate and regionality using a Bayesian Hierarchical Model?

2.1 Introduction

With increasing rates of Asthma throughout the USA, we thought that there was a correlation between asthma rates and how different regions of the USA are affected. To better understand this relationship, we also used the CDC Air Quality Indicator (AQI) data along with the CDC Chronic Disease Indicators (CDI) data. Using these data sets and the 4 regions from the previous research question (West, South, Northeast, Midwest), the continued use of the regions is mainly due to the minimal loss in generality along seen in Figure 1 and Figure 2 with the continuity with the research question in the previous section. However, we still do have to consider the outliers from the trend mentioned before on California. Mainly, the difference in PM 2.5 levels in the early 2000s compared to other Western states needs to be considered, but as we approach 2019, the AQI becomes more similar to the Western states.

In terms of Chronic Illnesses, we decided to group all the CDIs into 4 main groups: Asthma, Chronic Kidney Disease, Chronic Obstructive Pulmonary Disorder (COPD), and Cardiovascular Disease. Using these 4 categories, we had the option to chose one specific disease to further research and better understand. Overall, learning about the 4 different diseases, the ultimate chose came to be asthma data.

In this section, our team opted to approach the modelling of the disease to be continuous. Due to the temporal nature of the data, the use of time periods to be used as categorical variables along with the 4 regions created the framework for how we would approach our model. For a question like this, going into a Causal Inference approach also seemed to be difficult as the confounding variables and the complexity of the data. The confounding variables include but are not limited to State Healthcare Policies, Regional Income, and exposures to triggers. Along those lines, the use of multiple hypothesis testing seemed to simplify and detract from the model being created. Logically, the best next step seemed to be using temporal categories to predict and infer how asthma data will proceed in the future. By using a Bayesian Inference, the use of previous and current data can create a predictable model while new incoming data can be added on top to better estimate the predictions and distributions. The main modelling program being used are PyMC models to create prior and posterior distributions. Using this model, we will compute time as a categorical variable to simplify the data while not reducing the general complexity of the results.

2.2 Exploratory Data Analysis (EDA)

The main source of our data comes from the PM2.5 data as stated above, and the granularity of the data is annual PM2.5 level for each state. In terms of the CDI data, the granularity of the data before manipulation is annual type of disease levels for each state and question. From there, data was grouped into the 4 Major Chronic Diseases stated above, and each data point was categorized into the respective data sets for Asthma, COPD, Chronic Kidney Disease, and Cardiovascular Disease. The data was then cleaned to remove duplicates, extract data types as ratios, and remove NaN values. Then the data set was grouped by year, location, topic, and question and then filtered to contain each of the 4 Chronic Diseases. This reduced the granularity to annual disease levels for each state and question.

2.2.1 Asthma Data Analysis

To start things off, functions were defined using the Plotly Choloropleth map containing a map of the United States and a color gradient for how much of the chronic disease is prevalent per state in Figure 14 and Figure 15. This showed us a clear distribution of the cases of Asthma present for each state. Comparing the 2 graphs, the use of mean data showed more variability and importance as it represents the data on how often people are showing symptoms of asthma. In Figure 14, we are able to visually conclude that for the average cases of asthma showing a trend, northern and eastern states had more cases of asthma averaged to their population. We were then able to conclude that each region of states had similar values, and would not have too large of an effect if grouped into the 4 regions of the US.

For the chronic indicators dataset, we additionally noticed that there was a lot of duplicate data, as the data had been split into stratification categories and different adjusted rates. We addressed this by taking the overall data for each data point (rather than demographic-specific ones), the adjusted rather than crude rates, and scaling data values to the same scale.

Using the information gained before, Figure 8 was created showing the average cases of asthma per region. The data from figure 14 shows that there is a section of Unknown values which we disregarded as they represent U.S. occupying territories outside of the 50 states due to its conflict with the regional data. However, among the 4 regions, the mean asthma cases which ranged from 33 to 38. This allowed us to assume that the 4 regions have enough variability in the data to create Bayesian assumptions that would differ from each other.

Finally, to look deeper into what the data points mean and if they can be grouped in together without conflict, Figure 16 was made. This figure showed that a majority of the data was related to the prevalence and severity of asthma. This allowed us to assume that the data points presented can be considered to be used as the severity of asthma in these regions.

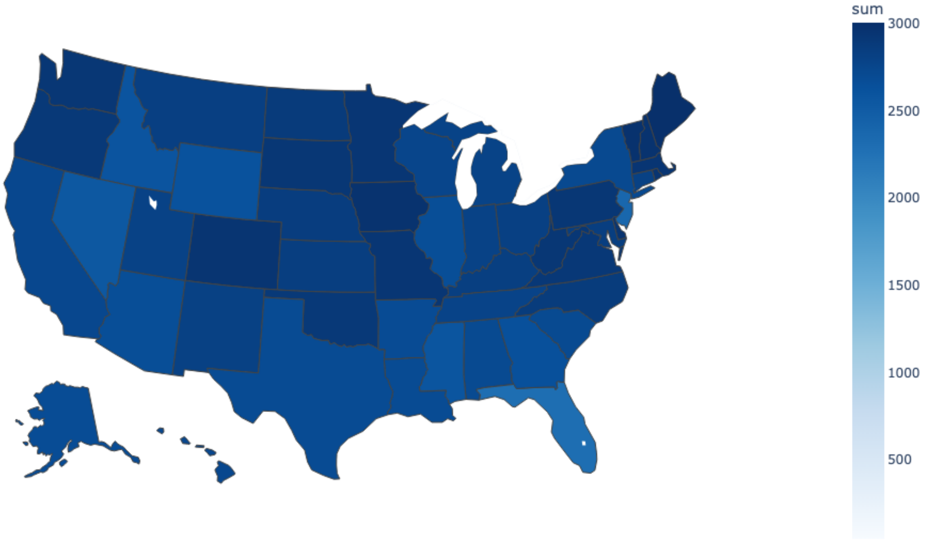


Figure 13: Count of Asthma Cases per US State

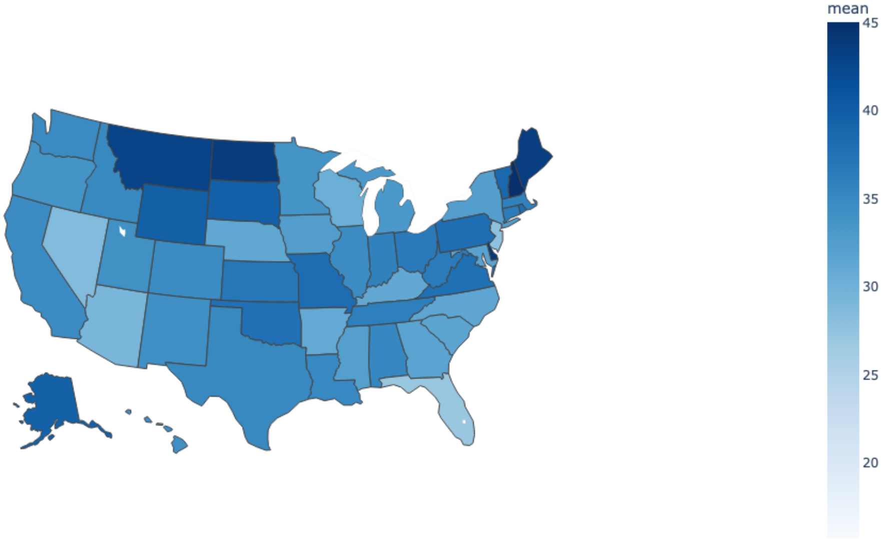


Figure 14: Mean of Asthma Cases per US State

2.2.2 Cardiovascular Disease, Chronic Kidney Disease, and COPD Data Analysis

Using similar functions used in the Asthma EDA to get the Gradient graph of the USA states and mean averages of disease per region, we found that asthma had the largest statistical variation of the mean as well as correlation to air quality. Not only that, when it came to questions, there was less variation in types of disease with asthma allowing for a more precise prediction using a model where grouping multiple different diseases into one category is not necessary. Therefore, the methodology to create a Bayesian Inference with Asthma was a unanimous and most logical decision moving forward.

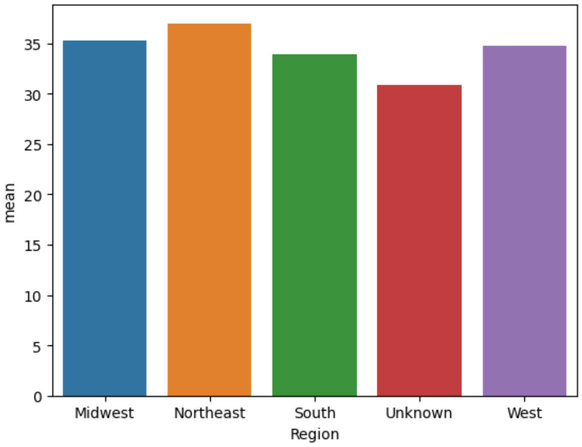


Figure 15: Distribution of Asthma for each US Region

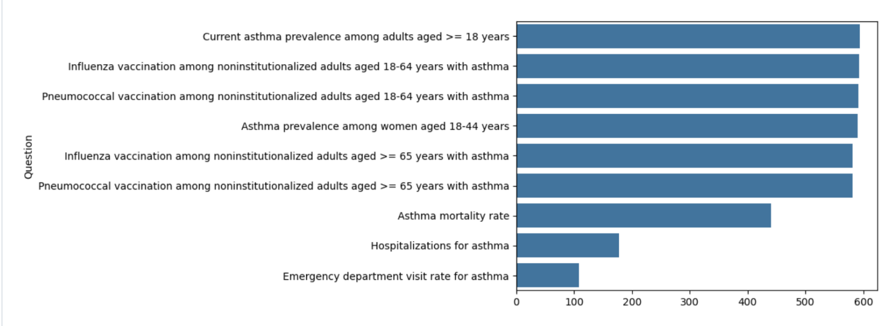


Figure 16: Distribution Questions in Asthma Data

2.3 Methods

2.3.1 Bayesian Hierarchical Modeling

Our goal is to find the relationship between asthma and the 4 regions in the US through Bayesian Hierarchical Modeling, particularly trying to determine the average in each region. First, we construct the naive estimates as the baseline, which we plot below. Next, in order to choose an informative prior, we plot the distribution of asthma per region (for a total of four plots), and note that each follows a normal distribution. The temporal data has been separated into 3 categories, before 2013, between 2013 to 2017, and after 2017. This separation allowed for us to run a PyMC model on the 4 regions of the US on the 3 time periods.

Our methodology on the model is that the Year will affect the size of our Population which should affect the amount of Asthma present. This is true since, if you look at the states with the largest populations, there is a higher prevalence of Asthma. Not only that, the Year affects how many people exists allowing for more asthmatic people. As for PM2.5, our hypothesis is that there is a higher likelihood of increased asthma with a higher PM2.5. With more particulate matter in the air, there is a higher likelihood of it containing particles that could trigger asthmatic attacks and worsen asthma conditions depicted by Figure 19.

When it comes to defining our Prior Distributions:

The PM2.5 distributions for each of the 4 regions are show in Figure 17. As all 4 graphs resemble a Empirical Distribution with skews and outliers, the 4 plots seemed to exhibit characteristics of a Normal Distribution. If we combine all 4 plots, it is safe to assume that the data follows a Gaussian Normal Distribution.

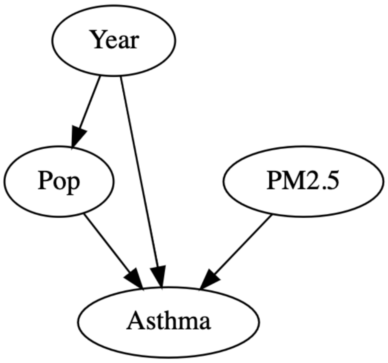


Figure 17: Graphical Model of Asthma

South and West Regions: These histograms show a right-skewed distribution, indicating that there are higher frequencies of lower PM2.5 values with some rarer occurrences of high PM2.5 values in these regions.

Northeast and Midwest Regions: The distribution appears more symmetric, especially for the Midwest, suggesting a more even spread of PM2.5 values around the mean.

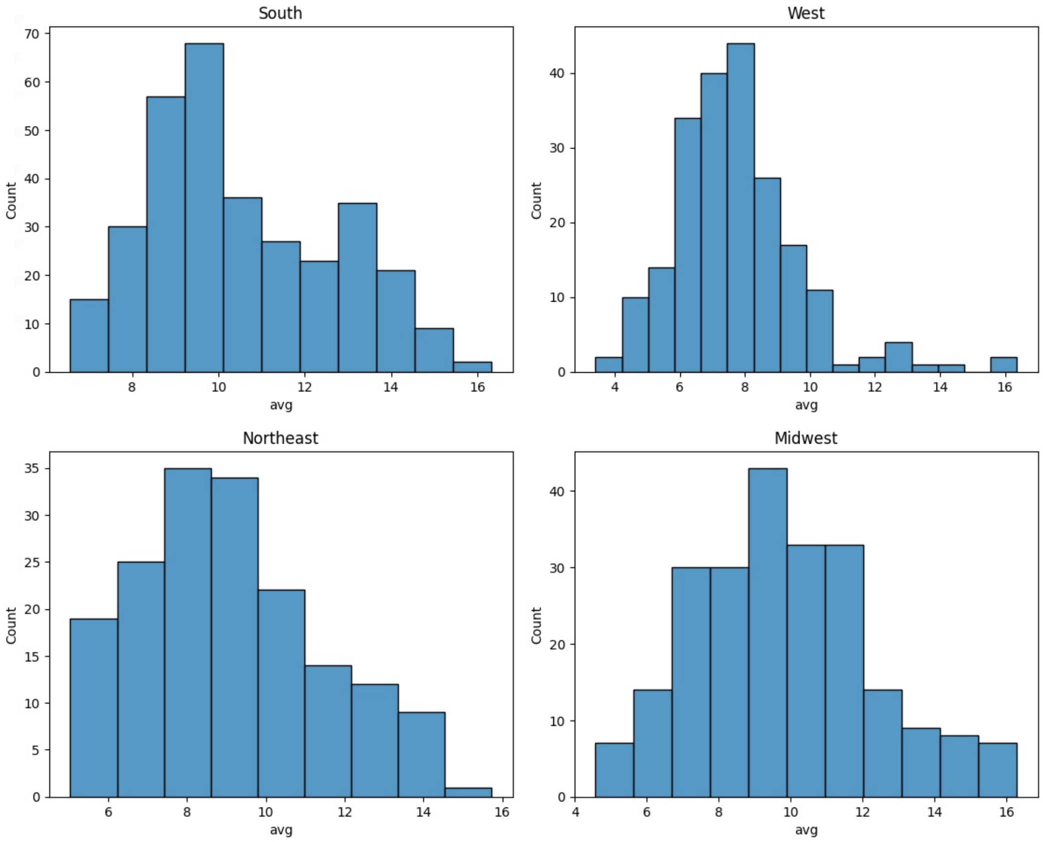


Figure 18: Prior Distributions for PM2.5

The asthma distribution is shown in Figure 18. The distributions seem to follow a Beta Distribution by looking at the graph below. When the graph of the 4 locations are put together, they seem to centralize and exhibit a Beta Distribution. The shape of the graph specifically looks like a Beta(2,5) in the Northeast, South, and West distributions with a bit of noise while the Midwest seems similar to a Beta(1, 1).

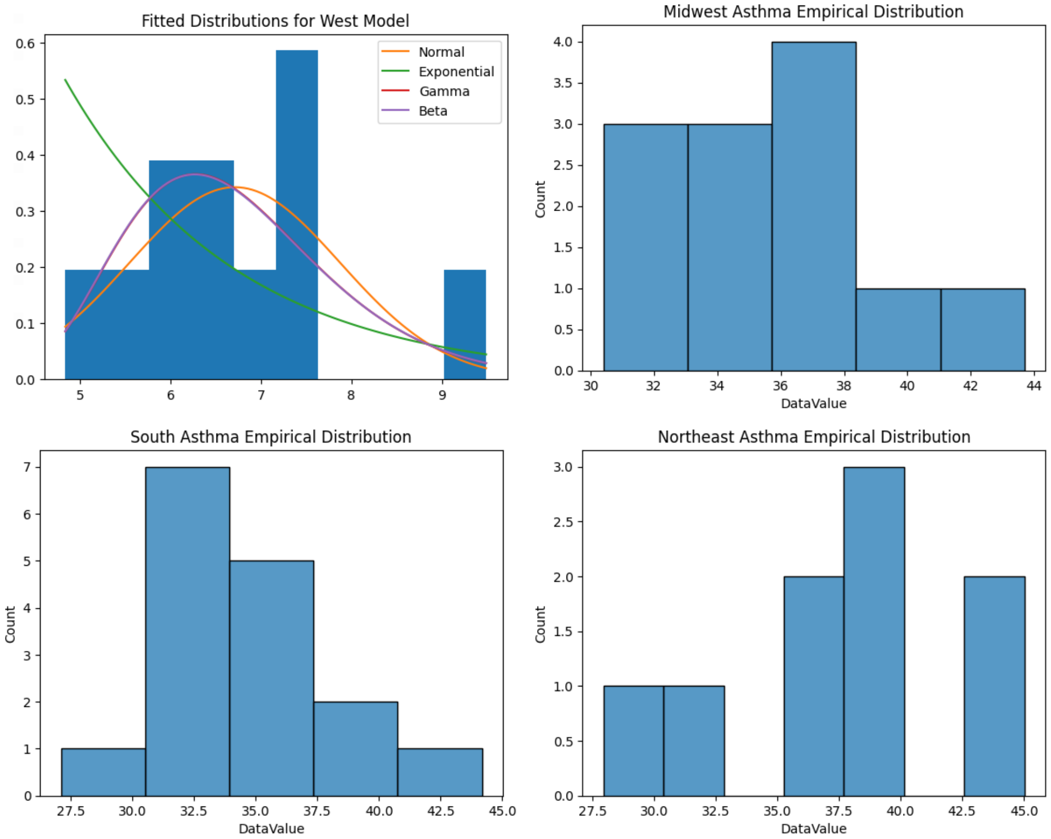


Figure 19: Prior Distributions for Asthma

The likelihood we used was a Beta Distribution that followed the Asthma graph to where the mean and variance were derived from each region and calculated with all our data.

2.4 Conclusion and Discussion

Through the use of PyMC and input of the previously mentioned prior distributions and likelihood, we were able to produce a posterior distribution and understand the conditional distribution of asthma based on differing regional and temporal factors. Figure 20 shows the Posterior Distribution compared to the actual values. Each Distribution follows the basic trend of the true distributions in a more normalized manner. Although the height of the distributions aren't as drastic, our model fairly predicts the rate of asthma all the regions pretty accurately except for the West Coast. It seems that the West has a large amount of asthma cases in the 30-35 percent range. Further, our analysis was severely limited to computational constraints, where only sampling a limited number of posterior distributions and limited to base distributions was possible. Had we been provided more computational resources, our posteriors would have likely reflected the true underlying distribution much more effectively.

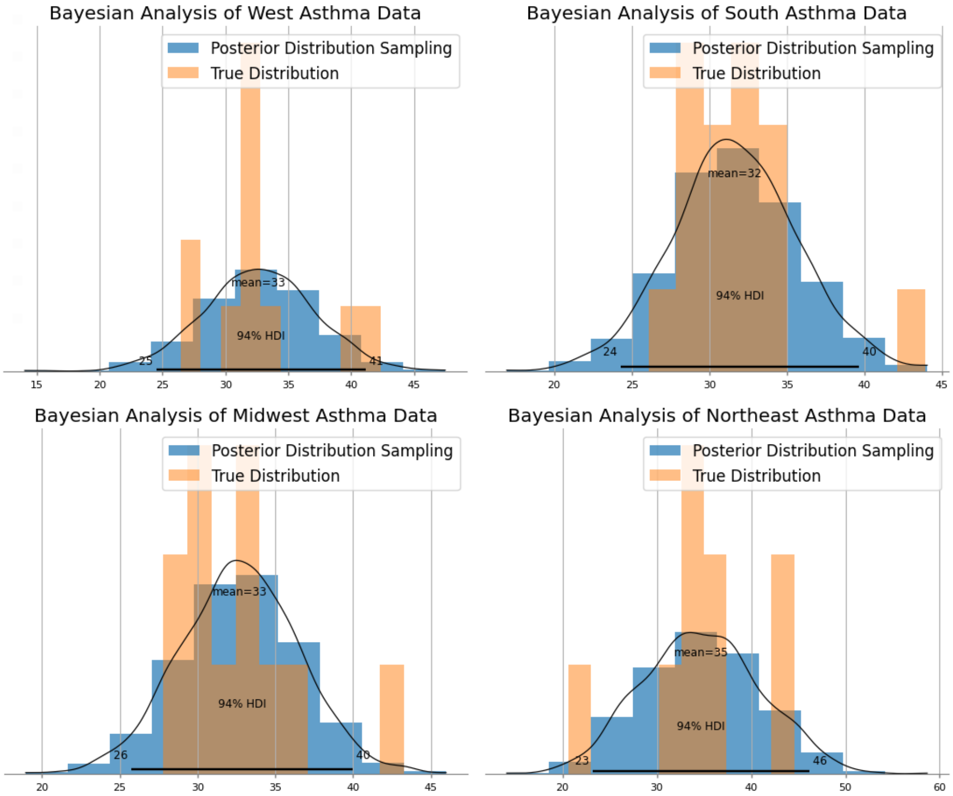


Figure 20: Posterior Distributions for Asthma

3 Conclusion

Given the results in our PM2.5 and asthma analysis, allocating resources to the southern region seems most feasible to improve overall health. By decreasing the PM2.5 levels to an area of the country must encumbered by asthma and air pollution, this results is determined most reasonable.

Our methods were limited to more recent data (2001 onwards), as well as the variables provided in the data sets we used, which may not have been very specific. We also limited ourselves to one specific disease (asthma) and AQI indicator (PM2.5) within the scope of this project. In the future, having more specific data highlighting more detail surrounding air quality and regional data could add to the Bayesian hierarchical model.

In combining the insights from both research questions, we conclude that the choice between ARIMA and GLM models is contingent on the stationarity of regional data, with ARIMA models excelling in non-stationary scenarios. The frequentist approach, particularly MLR and Ridge models, performed robustly across regions, offering interpretability and computational efficiency. Conversely, the

Bayesian Hierarchical Model illuminated the critical impact of regional and temporal factors on asthma prevalence, with the South identified as a priority for interventions due to high asthma and pollution levels, and increasing in both regards. Future investigations could benefit from more granular data, extending beyond asthma and PM2.5 levels to enhance model accuracy and policy relevance.