A Bayesian Analysis of Excess Deaths in the United States

Shrinidhi Rajesh
Ali Alirad
DePaul University
MAT 427

Introduction:

COVID-19 has had profound effects on the United States, with millions of confirmed cases and more than 1130000 deaths attributed to the virus. The examination of excess fatalities becomes a crucial tool in the never-ending quest to understand the full scope of the COVID-19 pandemic, providing a broader picture of the true magnitude of mortality. This comprehensive dataset navigates through weekly estimates of excess deaths, spanning various regions of the United States, thereby offering a wider perspective on the toll exacted by the pandemic. Beyond the confines of direct COVID-19 attributions, excess deaths are discerned as the discord between observed and expected mortality within specified temporal boundaries. The dataset not only covers deaths directly tied to COVID-19 but also sheds light on deaths from all causes, helping to uncover potential undercounting and understanding the complexities of mortality during these unprecedented times. To calculate extra deaths, the dataset uses strong Farrington surveillance algorithms, carefully considering variations in expected counts. While the estimates presented are provisional and adjusted for incomplete data, they are a valuable resource for understanding the many aspects of mortality during the ongoing pandemic.

This study takes on a big task—applying Bayesian statistics to estimate all-cause excess mortality in the United States over three years. Using a multiple regression framework, we aim to unravel the factors contributing to extra deaths, with the number of deaths as the focus. Using data from 2020 to 2021, our statistical approach seeks to find patterns, relationships, and nuances that shape mortality dynamics. Recognizing the complexity of modeling each state separately, we take a practical approach by grouping states into regions. This not only makes our analysis more scalable but also captures overall trends beyond state-specific differences. After making predictions for 2022 using the trained model, we critically evaluate the results to identify regions or states that closely match actual mortality patterns. This comparison gives us insights into the effectiveness of our predictions and reveals geographical differences in extra mortality during this specific time frame. As we delve into this analysis, the study contributes not only to the growing field of Bayesian statistics but also adds valuable knowledge to understanding how mortality is changing amid the ongoing challenges of the pandemic.

Preparation of Data:

The original CDC Dataset is a single excel spreadsheet from years 2020 to 2022. It compromises the total weighted and unweighted demographics of excess deaths from all causes and all causes excluding COVID-19, across each State and the District of Columbia. This dataset was first grouped by Year, its Type (Weighted/Unweighted), and then Outcome(All Causes/Excluding COVID-19), resulting in 9 distinct datasets each assigned a sheet in excel.

From here, the sheets were imported into R for further manipulation. States were filtered into 1 of 5 regions and four variables were retained: Week Ending Date, Observed Number of Excess Deaths, Upper Bound Threshold, and the Average Expected Count. The years 2020 and 2021 were grouped together as the training set to predict the observed number of excess deaths in 2022.

A weighted dataset of the excess mortality is used for the analysis. The assignment of different weights accounts for the varying contributions of different states. Additionally, the utilization of a weighted dataset in the analysis is crucial as it allows for a more accurate representation of the population and mitigates the impact of potential biases.

Variables
Week Ending Date:
The date marks the end of the respective week. Used to organize and categorize the data by week for a more granular analysis.
State:
The specific jurisdiction or state for which the data is reported. Used to categorize and analyze excess deaths on a state-by-state basis.
Region:
The 51 states are separated into 5 different regions. Used to categorize and analyze excess deaths based on these regions.
Observed Number:
The actual count of deaths observed during a specific week in a particular state. Provides real-time data on mortality, forming the basis for comparison with expected counts.
Upper Bound Threshold:
The higher limit for expected deaths. Used as thresholds to determine excess deaths. The difference between observed deaths and either of these thresholds helps identify the extent of excess mortality.
Average Expected Count:
The average expected number of deaths. Used as a baseline for comparison with observed mortality.

Exploratory Analysis:

Summary Statistics

Regions	Mean Of Observed Number	Range Of Observed Number	Variance Of Observed Number
Mid West	14502.3	7963	3544449
North East	10059.2	9810	3692144
South East	20382.41	9090	5652608
South West	7919.144	4910	1543290
West	11357.71	7893	2759974

Response Variable

Regions	Mean Of Average Expected Count	Range Of Average Expected Count	Variance Of Average Expected Count
Mid West	12711.36	1733	318444
North East	8881.298	1406	204925
South East	17373.8	2354	576890.8
South West	6446.106	942	90344.17
West	9768.788	1478	232570.8

Predictor Variable

Correlation

Within the correlation plots of each region, we run into our first issue with the dataset. While each of our independent variables has some degree of positive correlation with the dependent variable. The correlation among the independent variables across each Region is effectively perfect. As shown on the correlation plots, the Upper Bound Threshold and Average Expected Count has correlation value of 1. There is no reason to include both variables within our model. We have chosen to retain the Average Expected Count and drop the Upper Bound Threshold from our model.

Histograms

The histograms across the 5 Regions shows a consistent right skew for the Observed Number of Excess Deaths. Within linear regression, transformations are normally applied to variables that have skewed distributions. In this case, applying a natural logarithm to each variable does not eliminate or reduce its skewness. Our goal is to make a parsimonious model, therefore applying transformations that do not help normalize a distribution is counterintuitive.

The Average Expected Count of Excess Deaths shows a relatively flat distribution with noticeable spikes in tails across the Regions. The flatness in the distributions is expected since this data is weighted to account for missing data, as well as being a mean calculated value.

Plots

Average Expected Count and Observed count are plotted for each Region. These plots depict an atypical spread of the data, but some similarities with adjacent regions. The Southwest and Southeast show an almost "Butterfly" pattern in the plot. This indicates an underestimate of excess death at both and high value. The Midwest and West Regions depict underestimates at the mid-to-large number estimates of excess deaths while the Northeast Region's underestimates are in the center. Best of fit lines are superimposed on the plots that foreshadow large errors to come with our model.

Intricacies

The dataset initially included information on mortality for all causes, along with additional data excluding mortality. Upon comparing the datasets and calculating summary statistics, we observed a noteworthy finding. Excluding COVID-19 from the dataset, in essence, was nearly equivalent to including it. This might seem counterintuitive, but it stems from the fact that the dataset focused on excess deaths during the pandemic.

In cases where COVID-19 was excluded, the cause of death might be attributed to specific conditions like a heart attack. However, closer inspection revealed that the underlying cause of these deaths often traced back to the influence of the COVID virus. In other words, although the declared cause might be a heart attack, the root cause was linked to COVID-19. Consequently, there was little distinction between including and excluding COVID-19 in our analysis.

To streamline our focus and avoid redundancy, we chose to proceed with the dataset encompassing all causes. This decision allows for a more comprehensive analysis without duplicating efforts and enhances the clarity of our exploration into excess mortality during the pandemic.

Methodology

We employed the JAGS (Just Another Gibbs Sampler) framework to model our pre-processed dataset. Simple linear regression was utilized for subsequent analysis. Given a high correlation between two predictors, a deliberate choice was made to include only one predictor in the modeling process. The dataset was then stratified into a training set, comprising data from the years 2020 and 2021, and a distinct test set,

representing the year 2022. A region-specific approach was adopted, conducting Bayesian simple linear regression separately for each region. Priors for the regression model were chosen to adhere to a flat distribution, reflecting unknown prior assumptions. The model underwent 5000 iterations to ensure convergence and stability. The efficacy of the trained model was evaluated on the test dataset, specifically aiming to predict the observed number of death rates for the year 2022. This comprehensive methodology aims to provide transparency in our analytical process and forms the basis for subsequent interpretation and analysis of the model's predictions.

Fitting a Bayesian Model

The Model

Using only one independent variable, our model is: $\boldsymbol{Observed_i \sim Normal(\beta_1 + \beta_2(AEC)_i, \sigma^2)}$, for each $\boldsymbol{i}$ = ***Region***. Uninformative priors were used for each model: $\boldsymbol{\beta_1, \beta_2 \sim Normal(0, 100^2)}$ and $\boldsymbol{\sigma^2 \sim InvGamma}$ $\boldsymbol{(0.01, 0.01)}$. The reasoning behind using uninformative priors is that we're only using two years' worth of data, that is heavily affected by the COVID-19, to create our model.

Convergence

Four Monte Carlo Markov Chains were run for each of the 5 models with 10,000 post burn-in samples before 15,000 iterations to establish convergence. Each parameter, β_1, β_2, converged within these iterations for each chain. Interestingly, each density plot for each Region shows that '0' is very close to the center of the distribution for their β_1 parameter, indicating that the model may not require a y-intercept.

Predictions

Summaries of each of the models shows consistency among the Region's β_1, β_2 parameters. Each Region's model was used to predict the 2022 Observed Excess Deaths. Among the 5 models, the West Region's Model has the lowest Mean Square Error while the Southeast Region has the highest. Due to the variance in our data being very high, our error statistics are high as well. Normalizing the data would be a recommendation when replicating the analysis.

Regions	MSE	RMSE	MAE
MidWest	2906498	1704.845	1497.989
North West	1311369	1145.15	996.8603
South East	5036022	2244.108	2035.432
South West	1150399	1072.566	990.4863
West	1129958	1062.995	892.321

Conclusion

This study began by analyzing a comprehensive dataset from 52 states via the CDC, offering a detailed view of COVID-19's impact on mortality. Through thorough data cleaning and preprocessing, we focused on three key columns and conducted regional analysis to better understand mortality patterns. Facing challenges like high correlation between predictor variables during multiple linear regression, we chose a singular variable and divided the dataset into training and test sets for Bayesian simple regression using JAGS across regions. Despite grappling with issues like variable correlation and skewed distributions, we aimed for a balanced model reflecting the nuanced dynamics of excess mortality.

After the analysis, our predictive model, though not meeting all expectations, highlighted regional disparities. The West region performed comparatively well, with lower errors, while the Southeast posed complexities in mortality prediction. This discrepancy emphasizes the intricacies in excess mortality estimation, urging ongoing refinement. This study contributes to Bayesian statistics but underscores the need for continuous improvement. Moving forward, exploring additional variables and refining models is crucial for enhanced predictive accuracy. The evolving pandemic demands adaptability, with methodologies honed to match the dynamic landscape of excess mortality.

By forecasting regions with higher expected mortality rates due to COVID-19, the model facilitates resource allocation, aiding healthcare authorities in distributing medical resources, prioritizing vaccination campaigns, and optimizing hospital capacities. Policymakers can utilize the predictions to inform region-specific interventions, adjusting policies such as lockdowns and travel restrictions. The model also assists in awareness and adherence to preventive measures. Furthermore, its adaptability allows for continuous monitoring and updates, ensuring that responses remain dynamic and aligned with the evolving nature of the pandemic.

This study, with its relative successes and challenges, provides valuable insights into excess mortality estimation. As the pandemic unfolds, it emphasizes the ongoing necessity for refining methodologies and models to truly understand COVID-19's impact on mortality across diverse U.S. regions.

Limitations

The study, while offering valuable insights into excess mortality prediction, has certain limitations that requires consideration. Firstly, reliance on the CDC dataset introduces potential data inaccuracies and biases that influences the robustness of the conclusions. Additionally, the pandemic dataset was replacing the missing values using Frimighams Algorithm. The effectiveness of this algorithm is questionable in replacing the missing values, as most of the dataset is unknown of its accuracy.

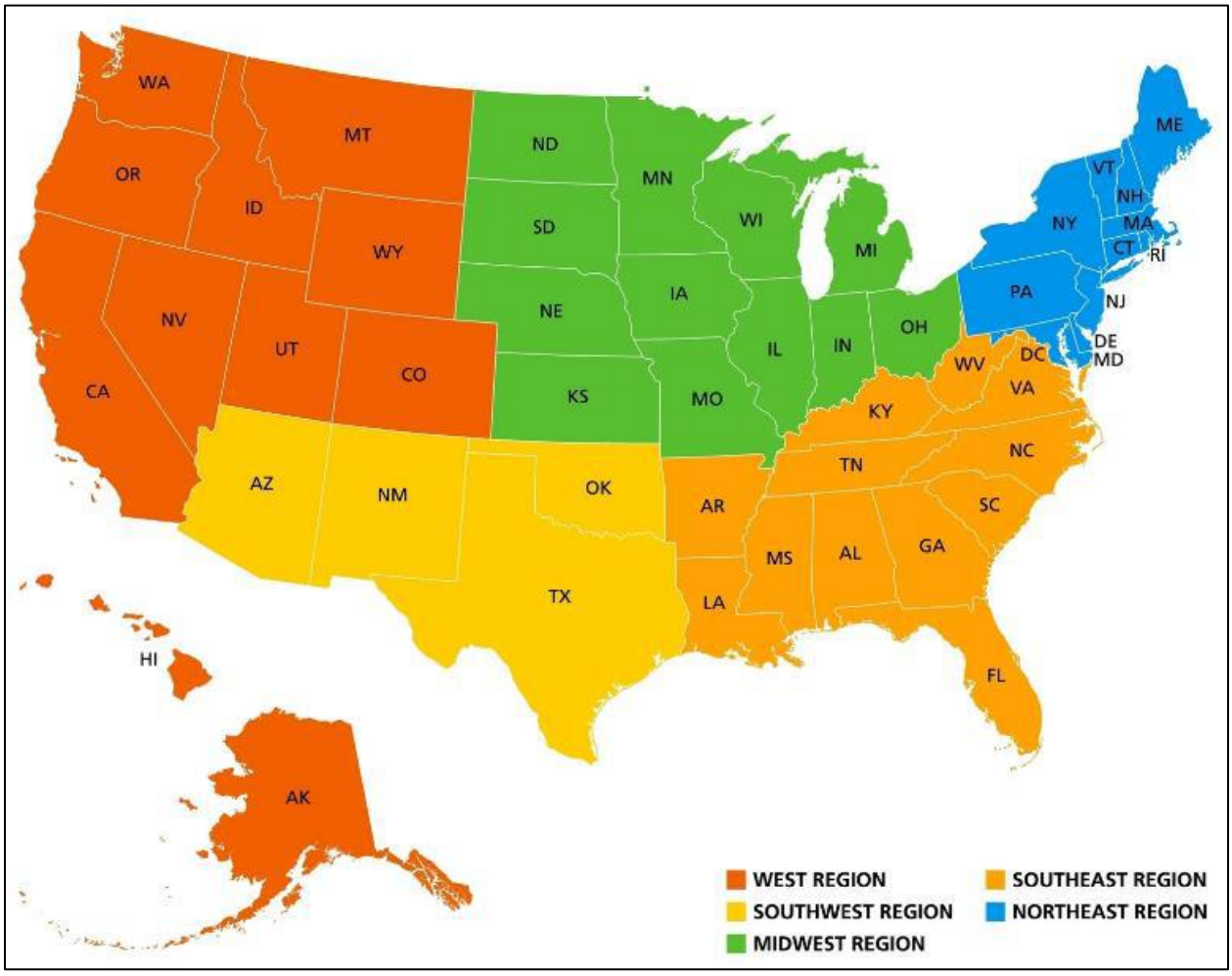
The chosen Bayesian simple regression model using JAGS, while insightful, may not fully capture the complexity of excess mortality dynamics, as evidenced by challenges like variable correlation and skewed distributions. Regional analysis, although providing an overall perspective, may oversimplify the diverse factors within each region that could impact mortality predictions. The study's focus on three key columns might overlook other pertinent variables influencing mortality, limiting the model's explanatory power.

Thus, cautious interpretation and further validation are essential when extrapolating the findings to broader contexts or timeframes.

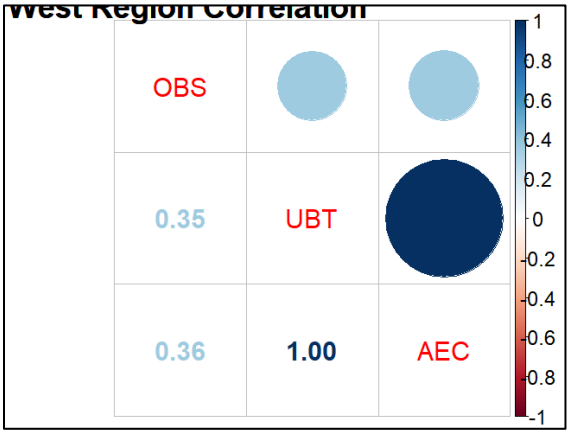
Recommendations

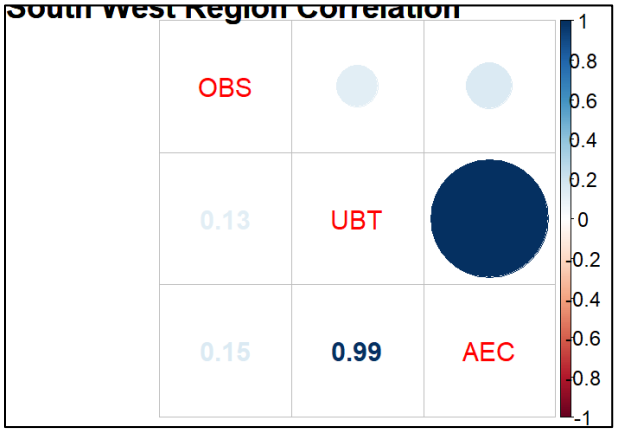
The Bayesian simple regression model using JAGS may benefit from refinement or alternative modeling approaches to better capture the complexities of excess mortality dynamics. Exploring advanced statistical techniques or machine learning methods could address challenges like variable correlation and skewed distributions more effectively. To perform advanced modeling a comprehensive validation and enhancement of the CDC dataset is recommended to address potential redundancies and biases. Verification through additional sources and scrutiny of the Framingham Algorithm for replacing missing values is essential to ensure the dataset's accuracy to perform further analysis. The focus on only two columns may be limiting, and incorporating additional relevant variables could provide a more understanding of the factors contributing to excess mortality. Furthermore, using more data across many years can help in developing and more informative priors to make a better model. A broader scope in terms of variables influencing mortality needs to be considered to enhance the model's explanatory power.

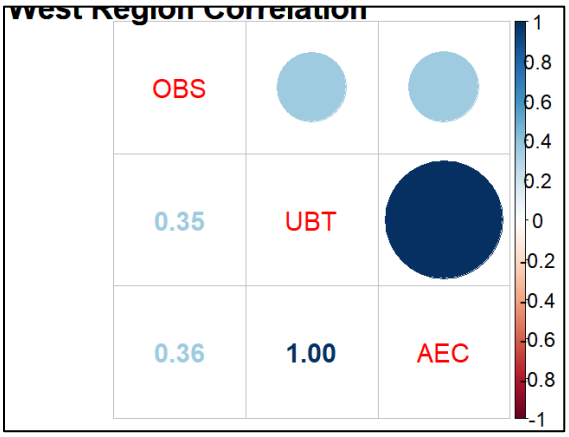
Regional Map

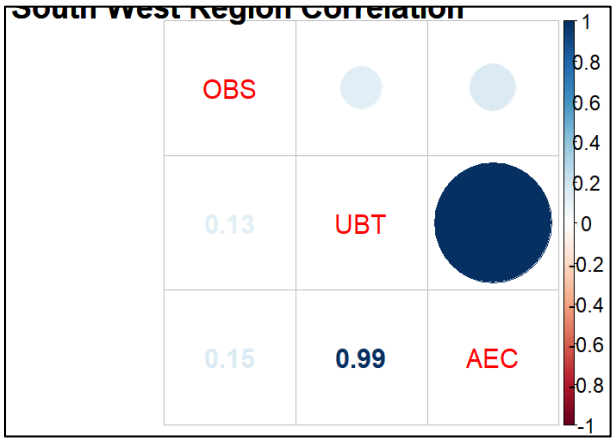


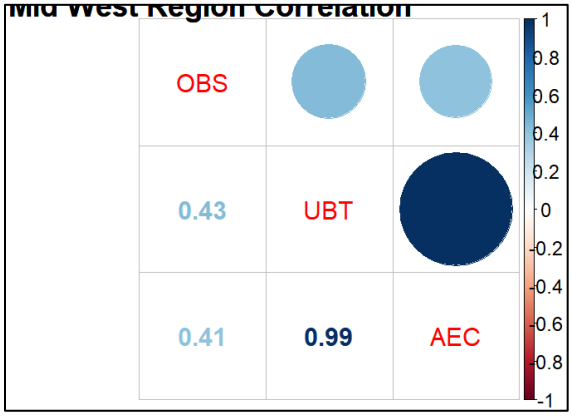
Correlation Matrices

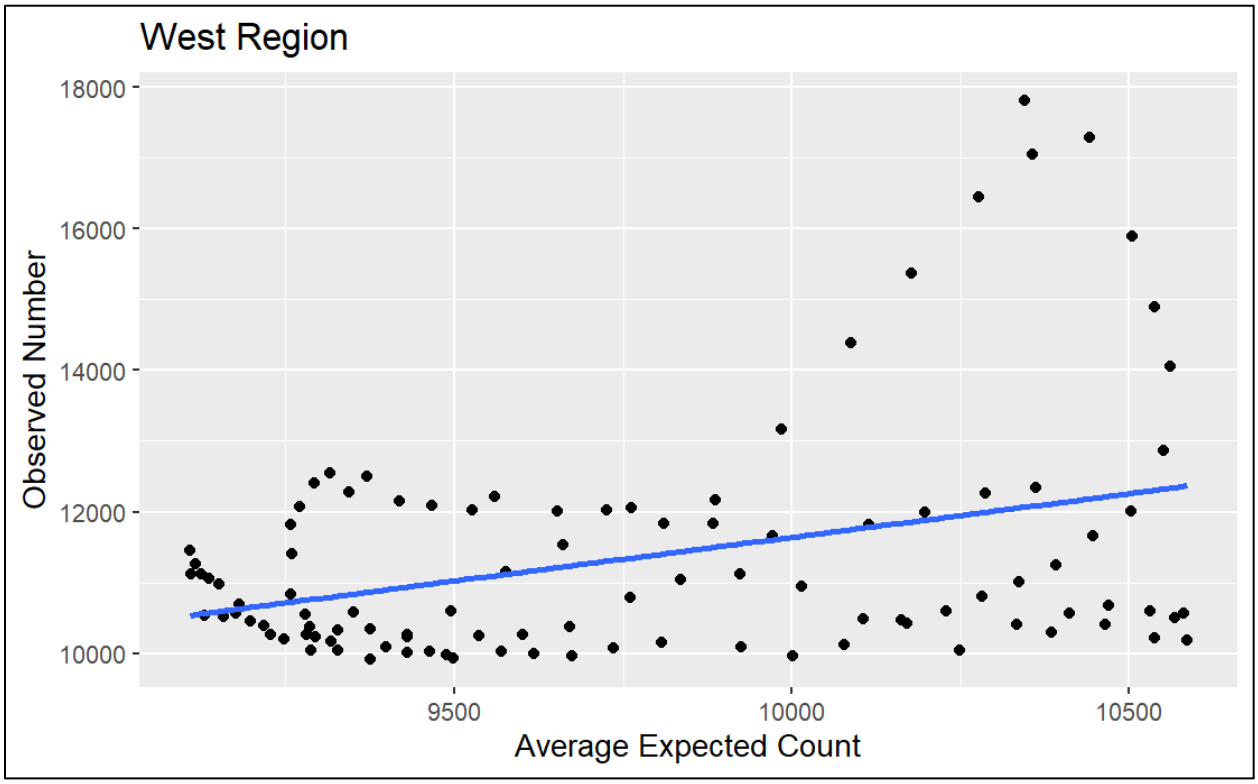


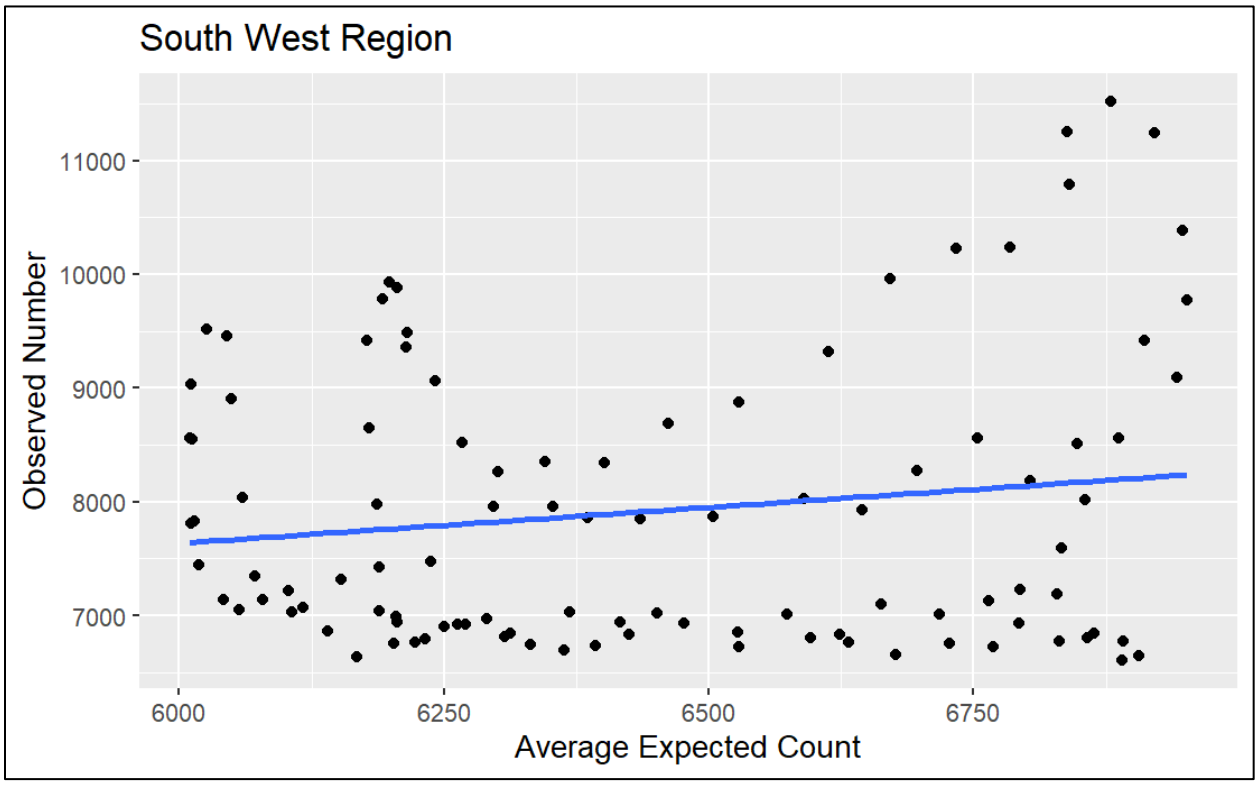


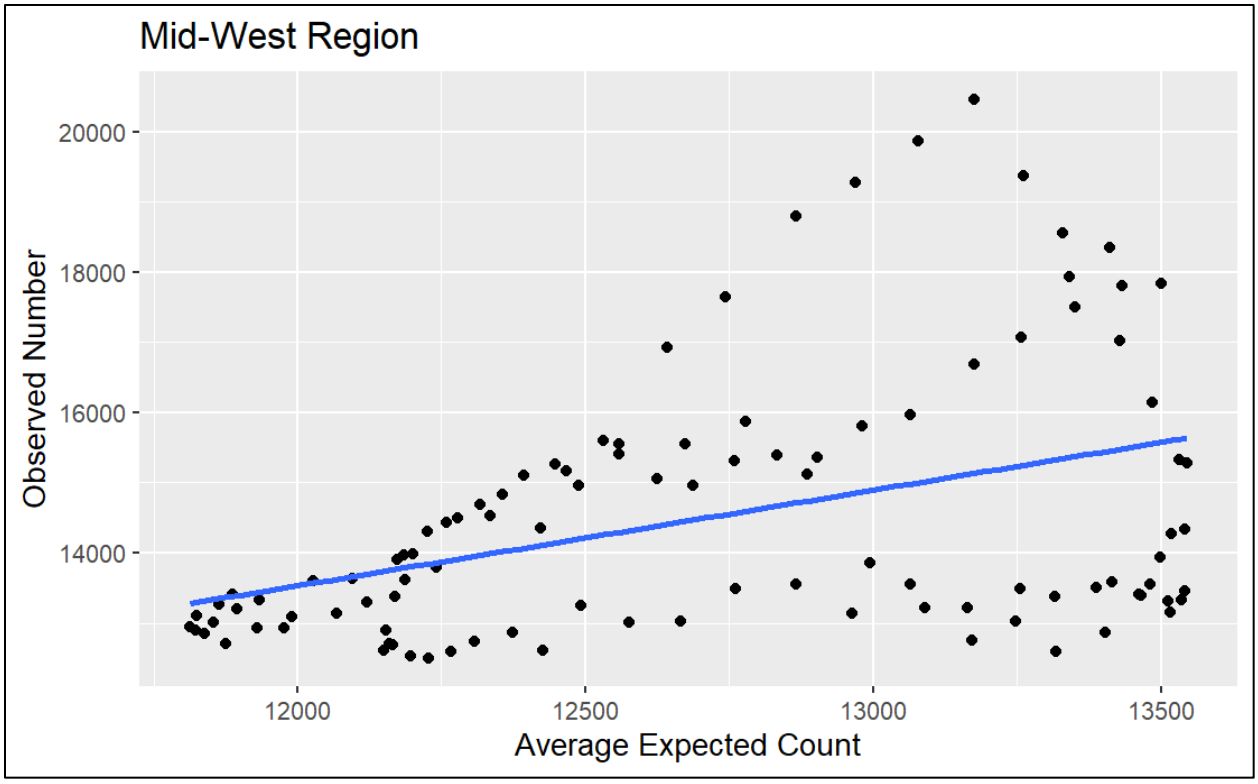


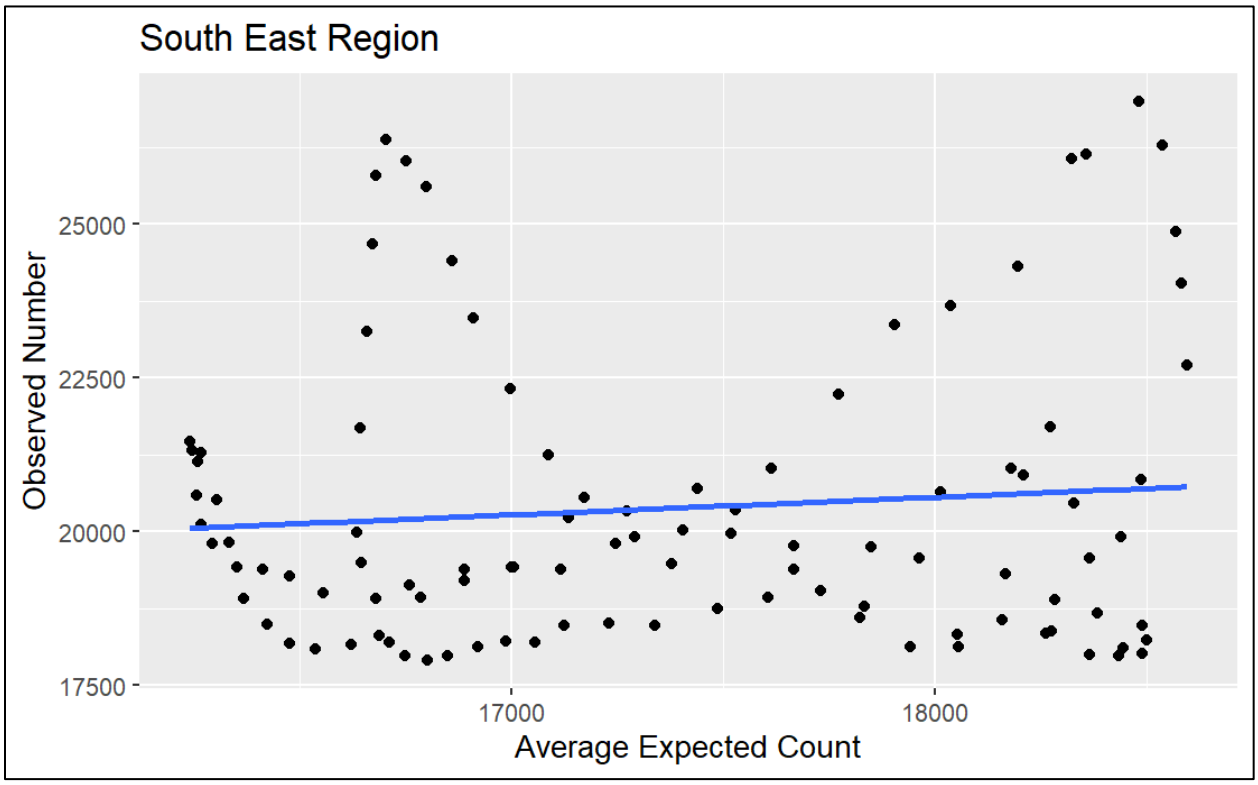


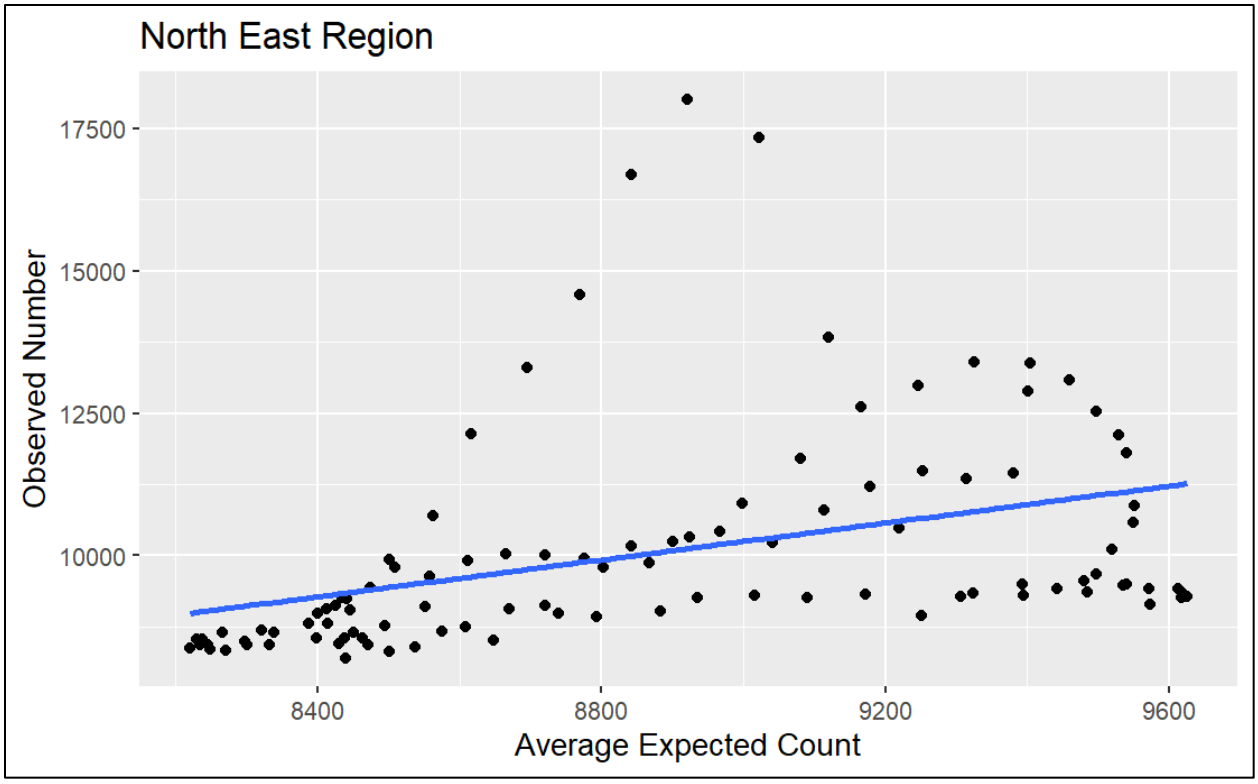


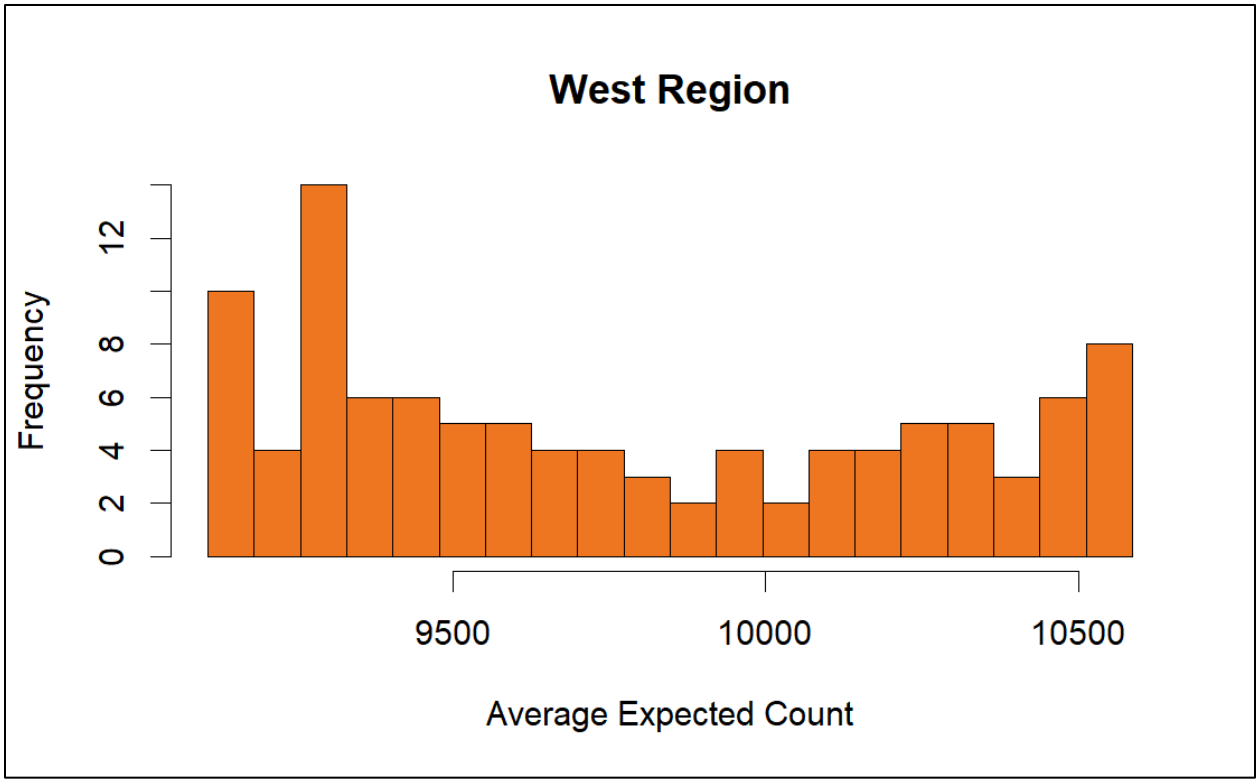

West Region
18000
16000
14000
12000
10000
Observed Number
9500
10000
10500
Average Expected Count

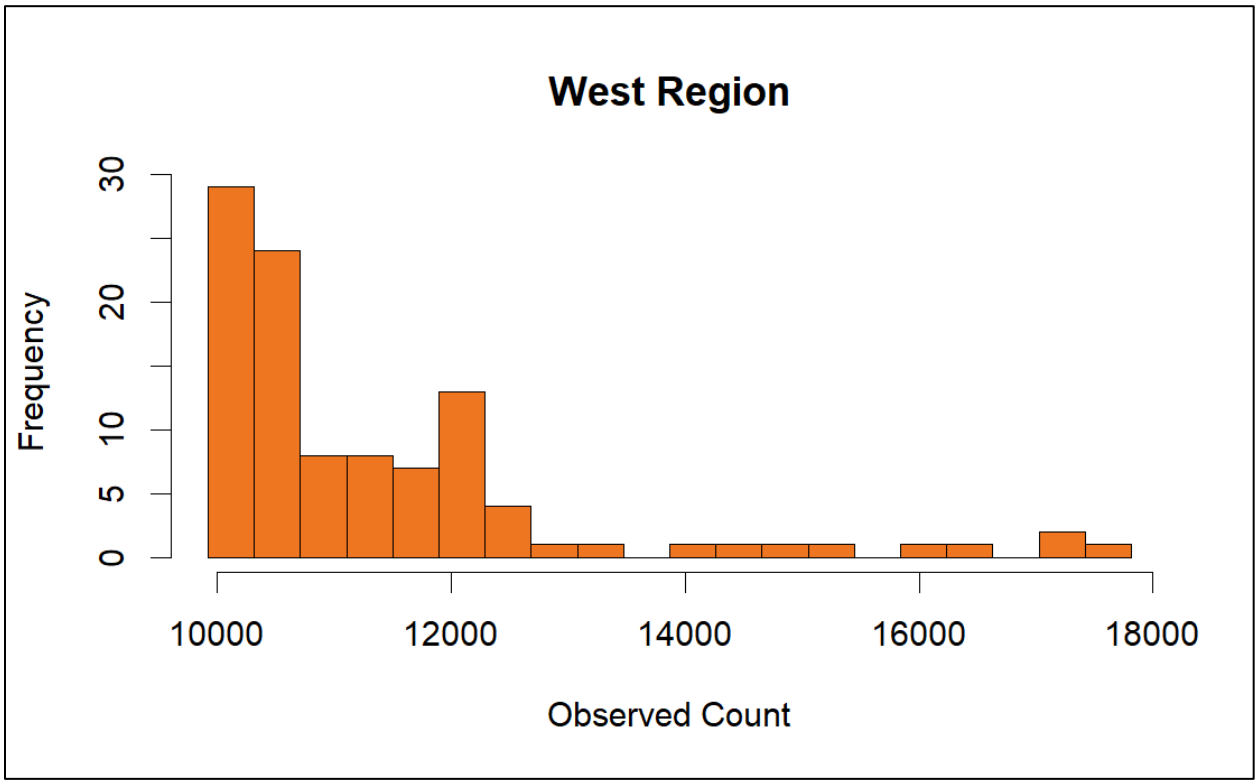

South West Region
11000
10000
9000
8000
7000
Observed Number
6000
6250
6500
6750
Average Expected Count

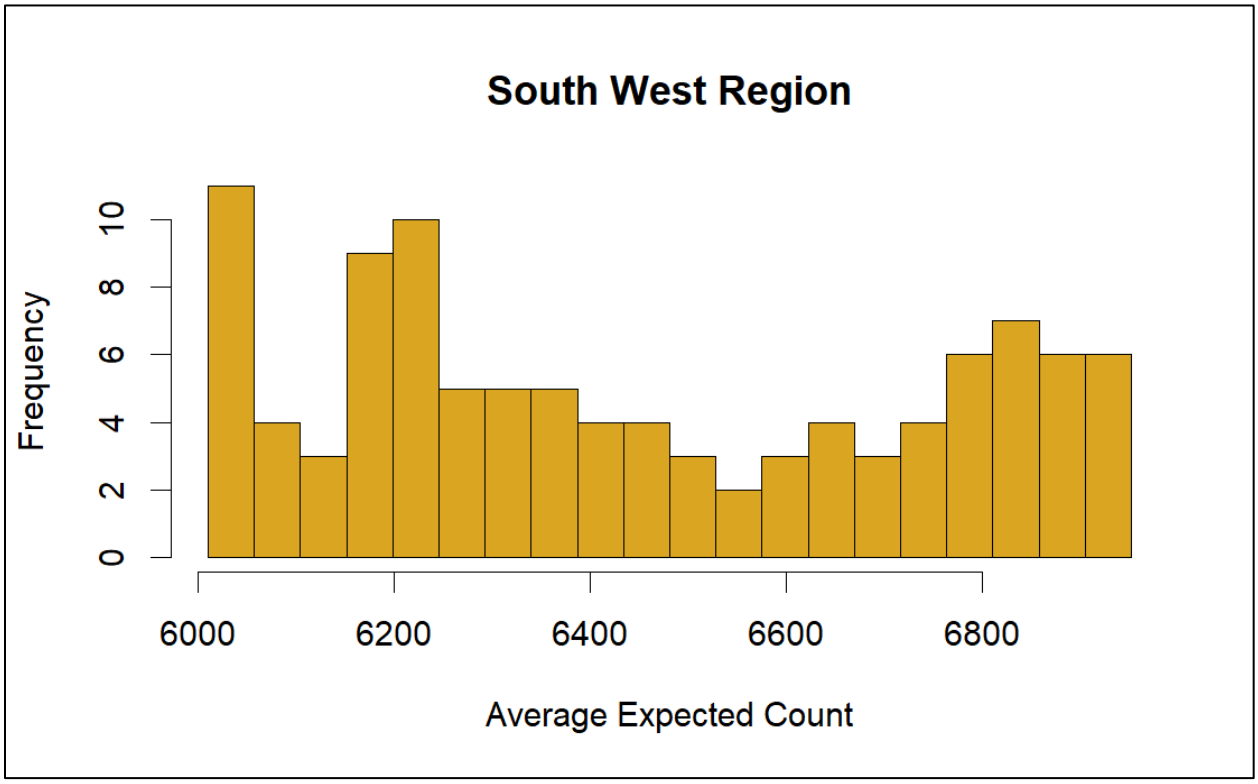

Mid-West Region
Observed Number
20000
18000
16000
14000
12000
12500
13000
13500
Average Expected Count

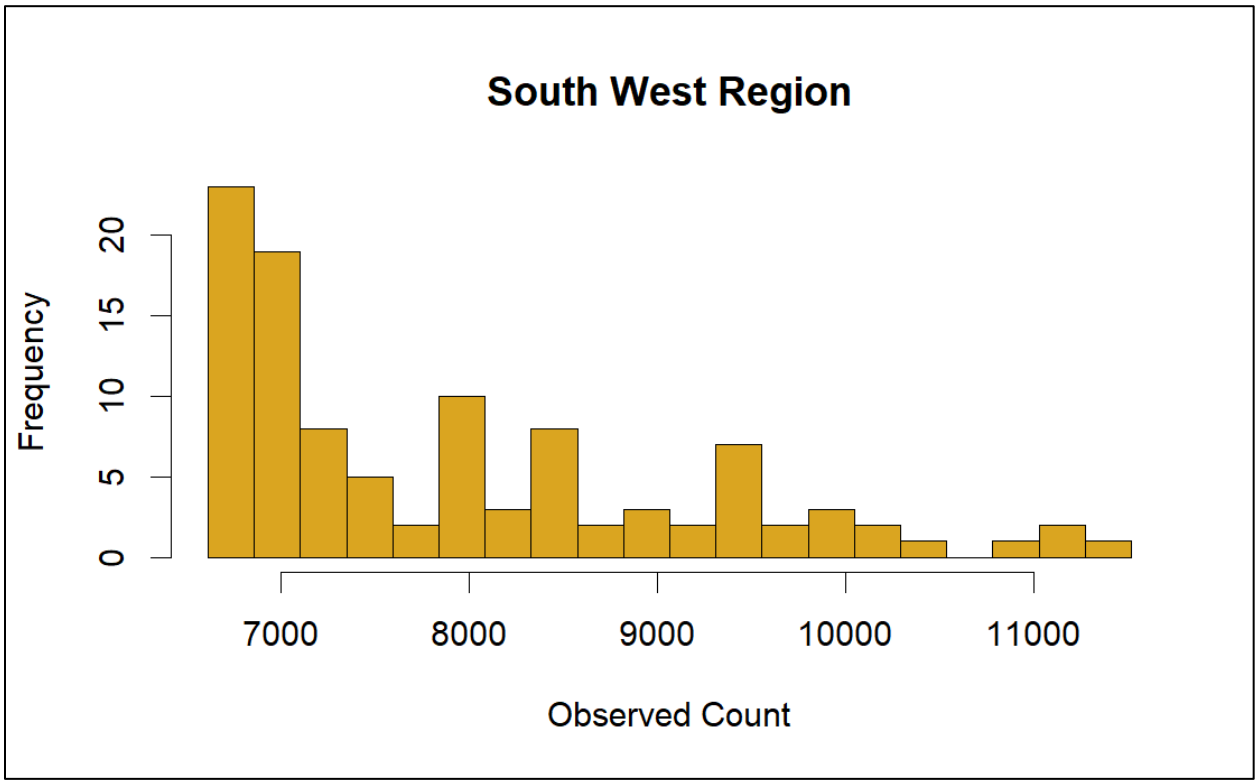

South East Region
Observed Number
25000
22500
20000
17500
17000
18000
Average Expected Count

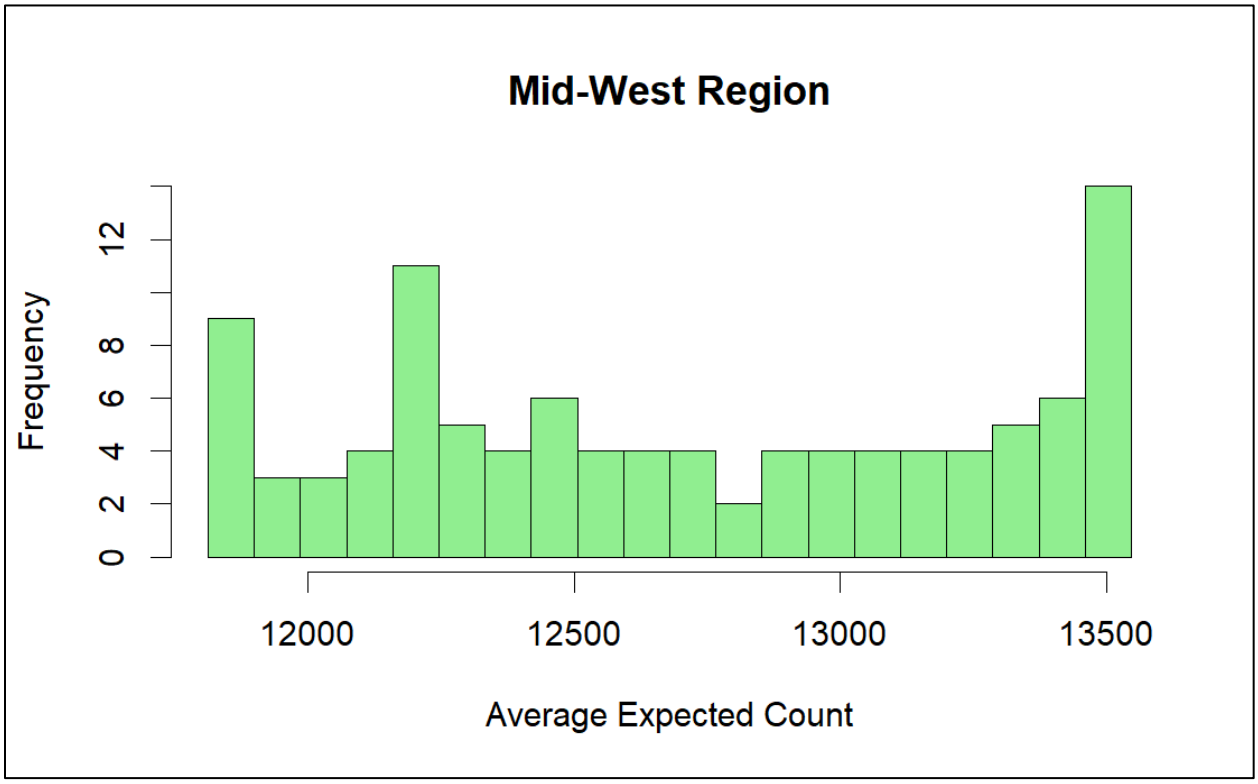

North East Region
Observed Number
17500
15000
12500
10000
8400
8800
9200
9600
Average Expected Count

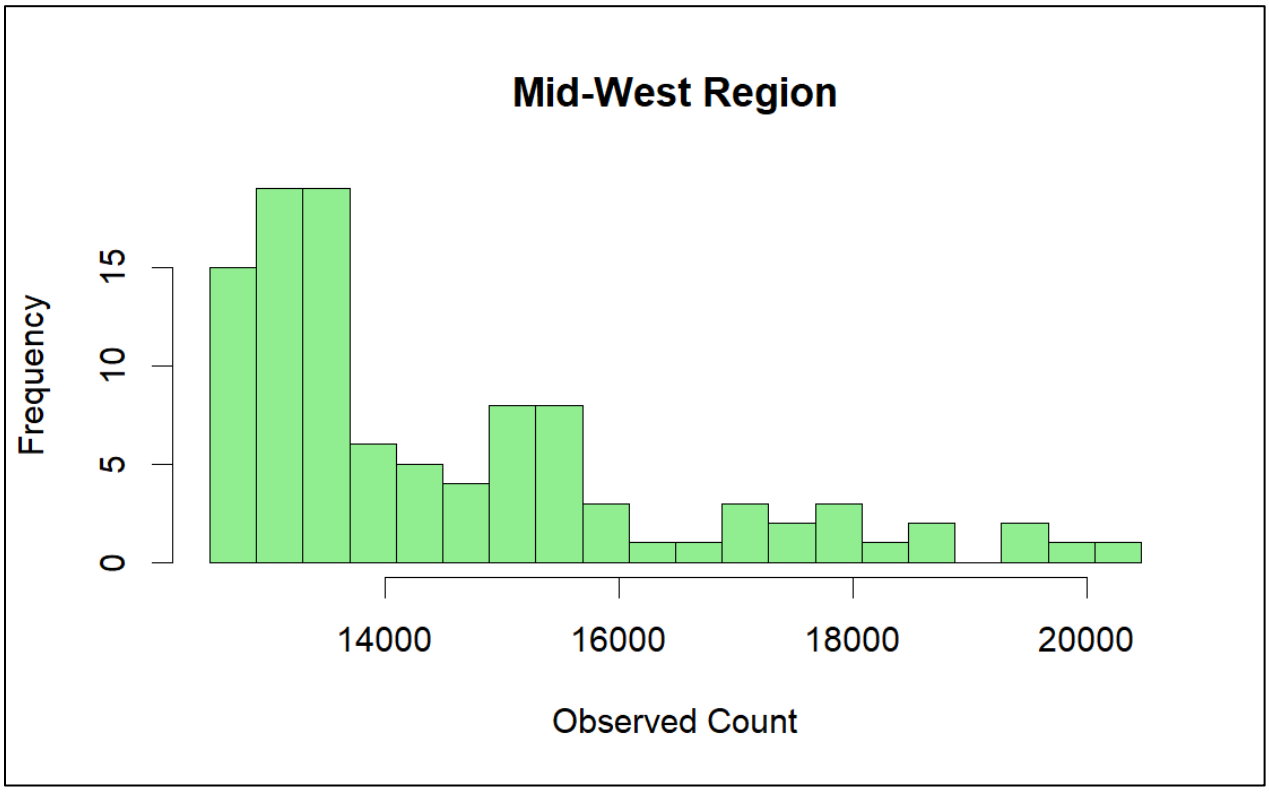

West Region
Frequency
0
2
4
6
8
12
9500
10000
10500
Average Expected Count

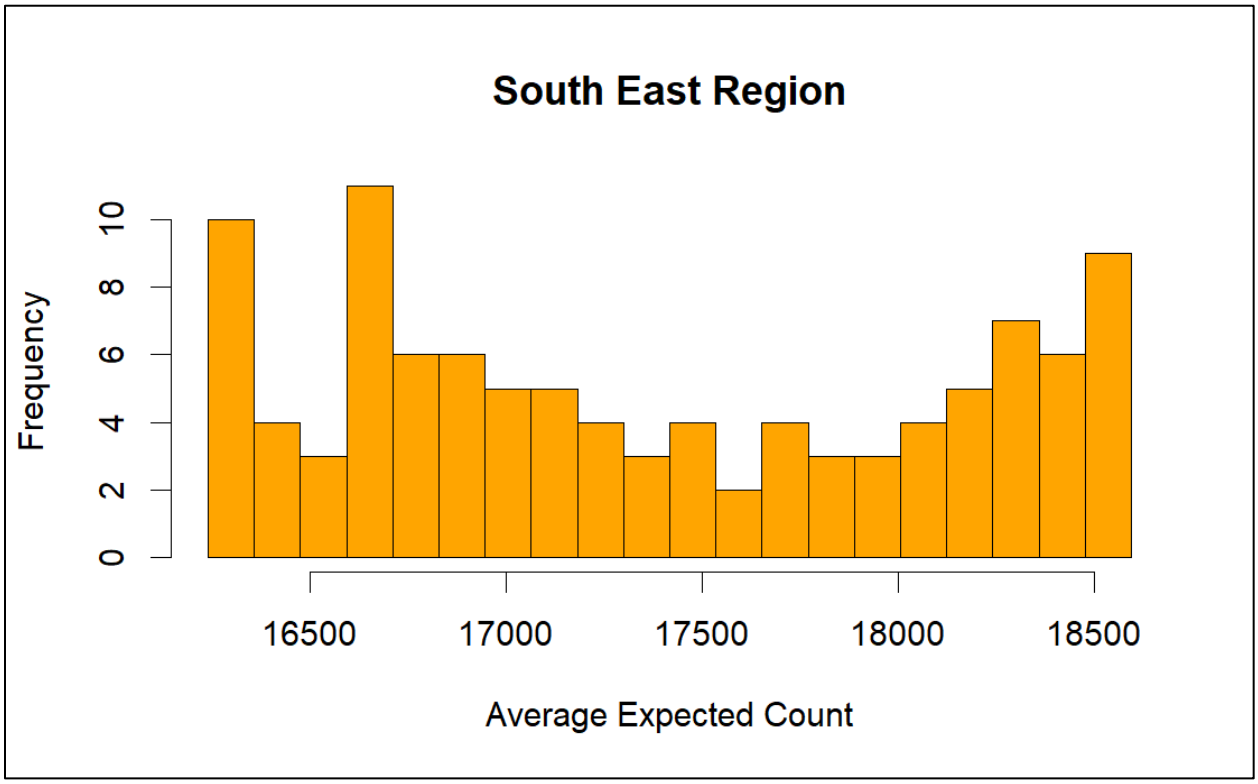

West Region
Frequency
0
5
10
20
30
10000
12000
14000
16000
18000
Observed Count

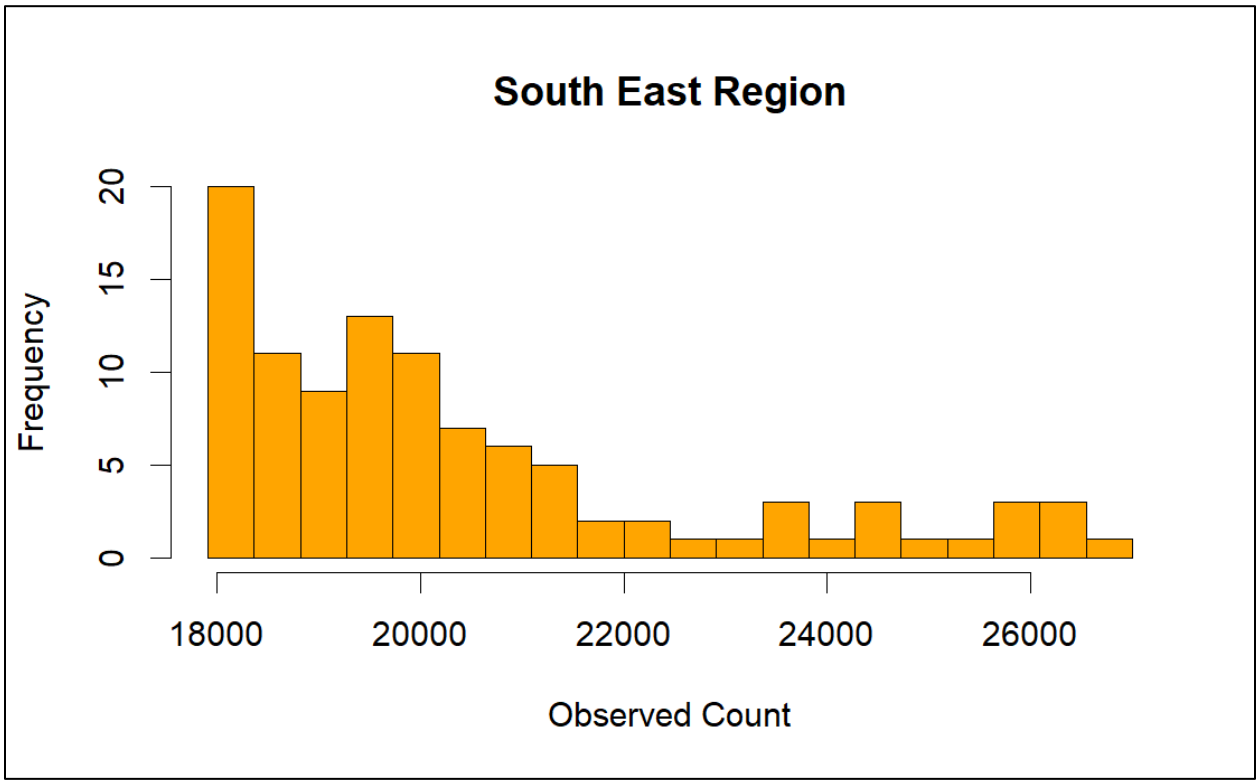

South West Region
Frequency
0
2
4
6
8
10
6000
6200
6400
6600
6800
Average Expected Count

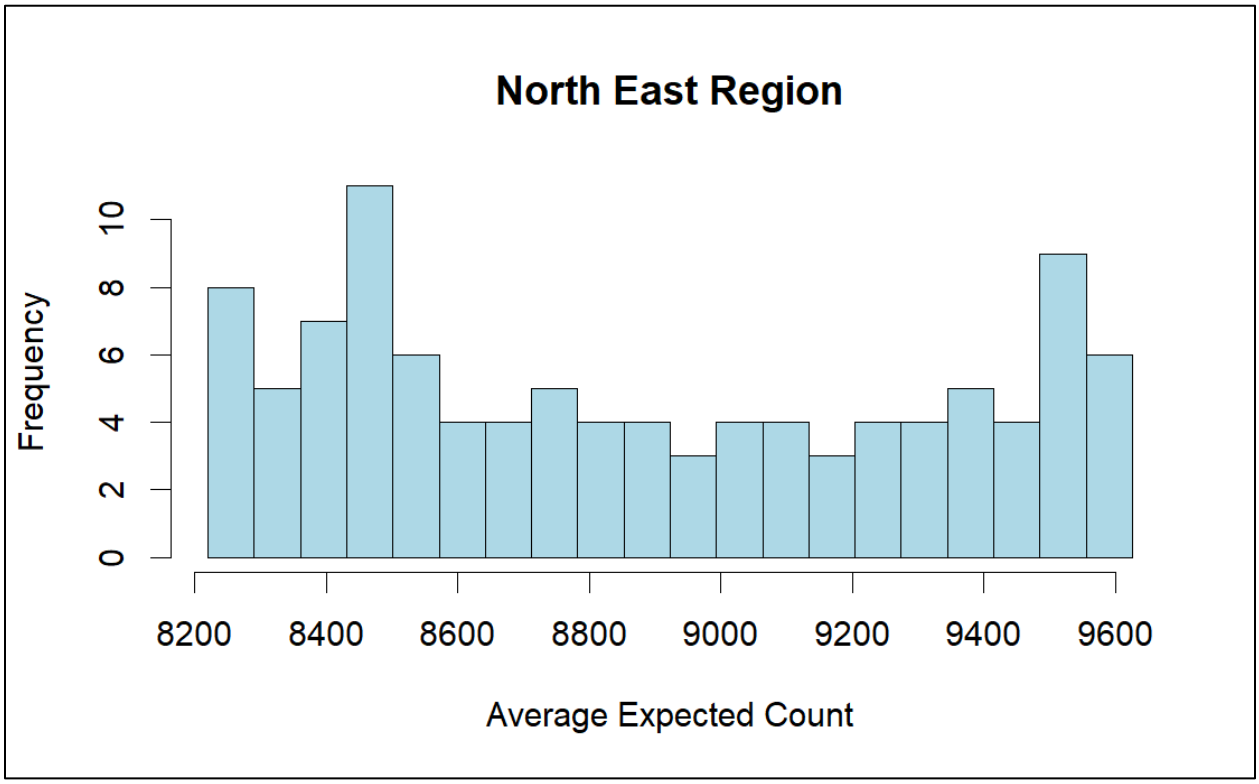

South West Region
Frequency
0
5
10
15
20
7000
8000
9000
10000
11000
Observed Count

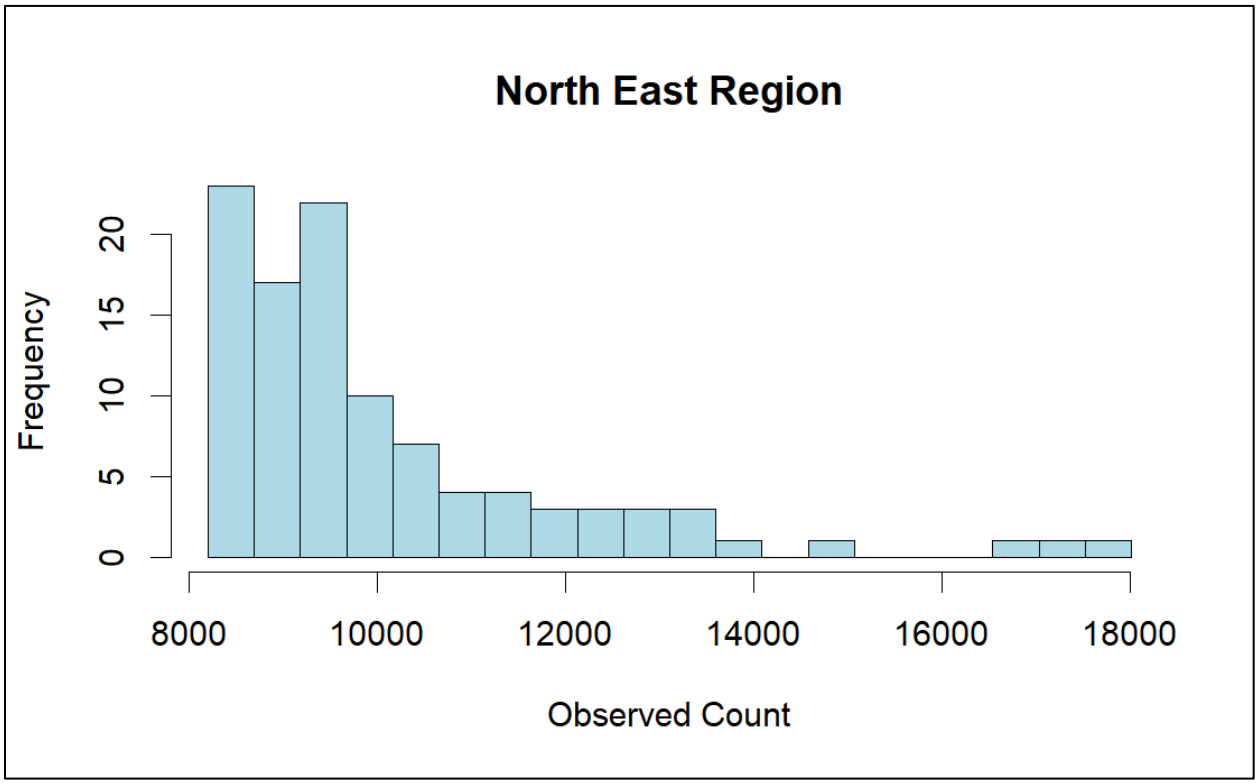

Mid-West Region
Frequency
0
2
4
6
8
12
12000
12500
13000
13500
Average Expected Count

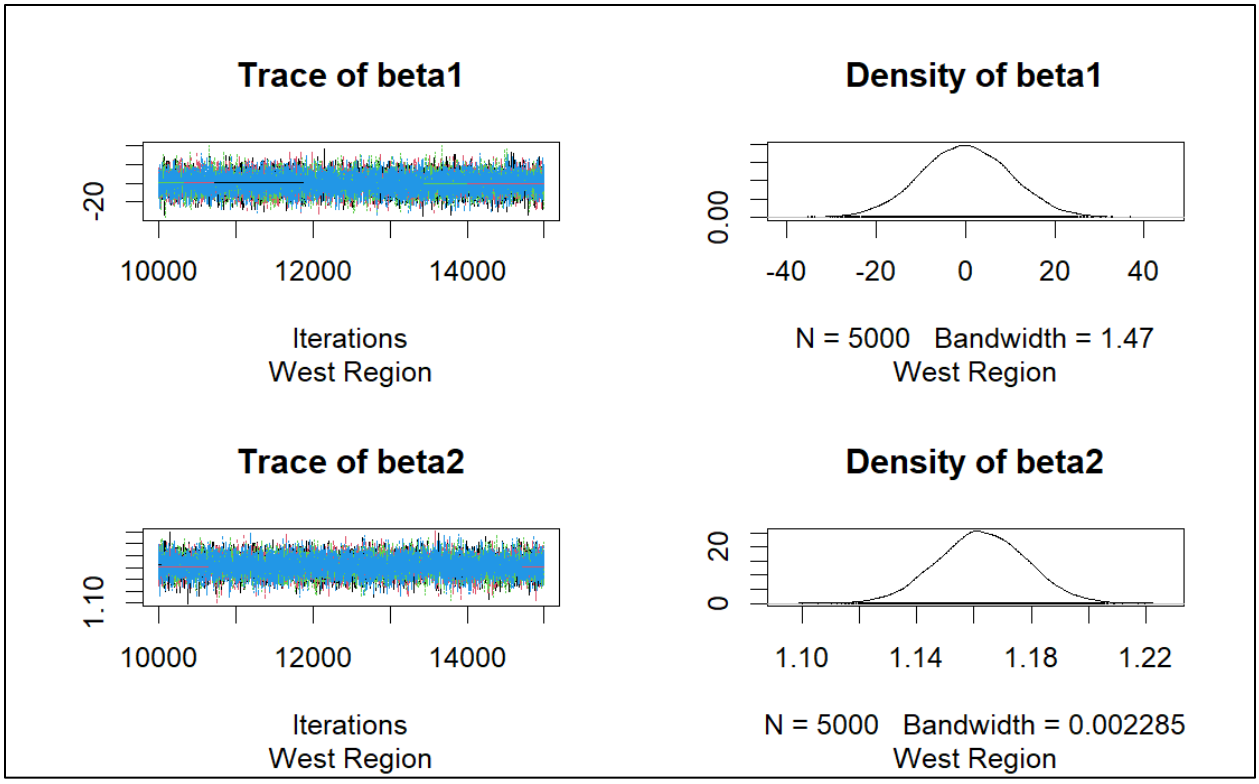

Mid-West Region
Frequency
0
5
10
15
14000
16000
18000
20000
Observed Count

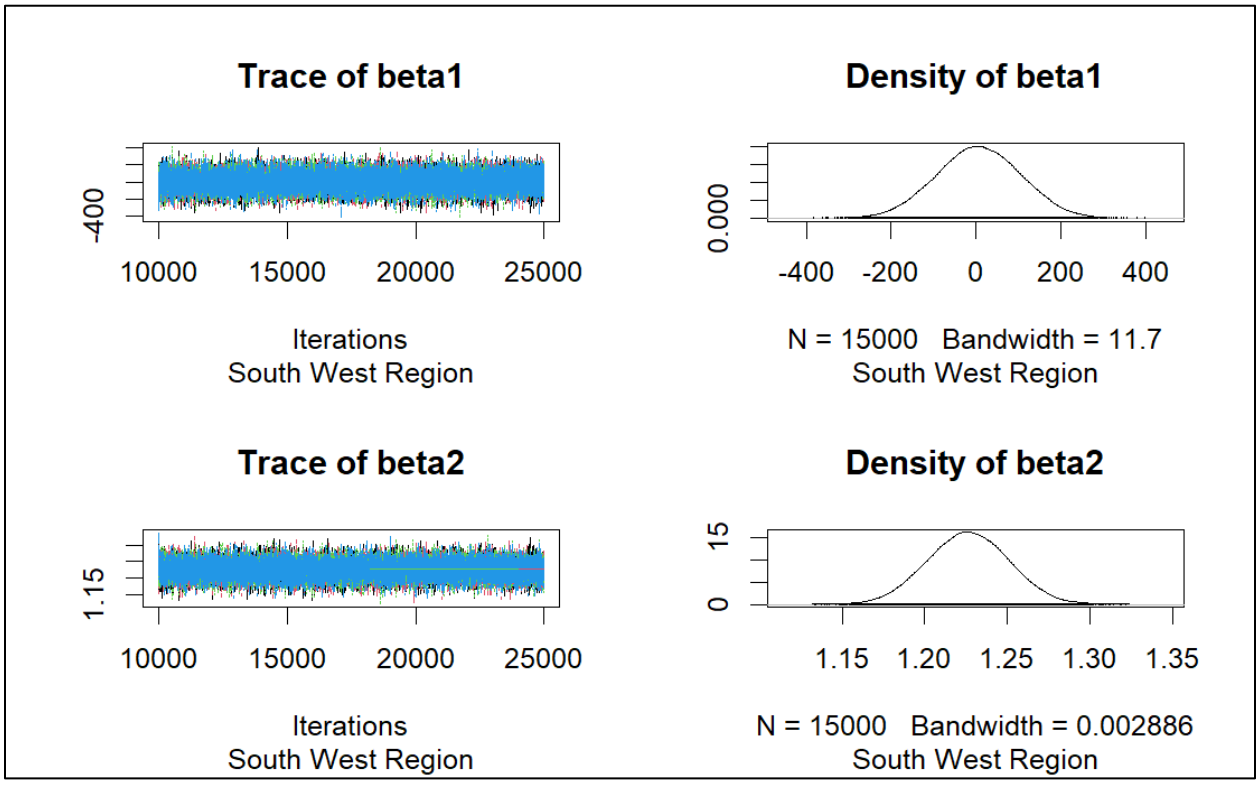

South East Region
Frequency
0
2
4
6
8
10
16500
17000
17500
18000
18500
Average Expected Count

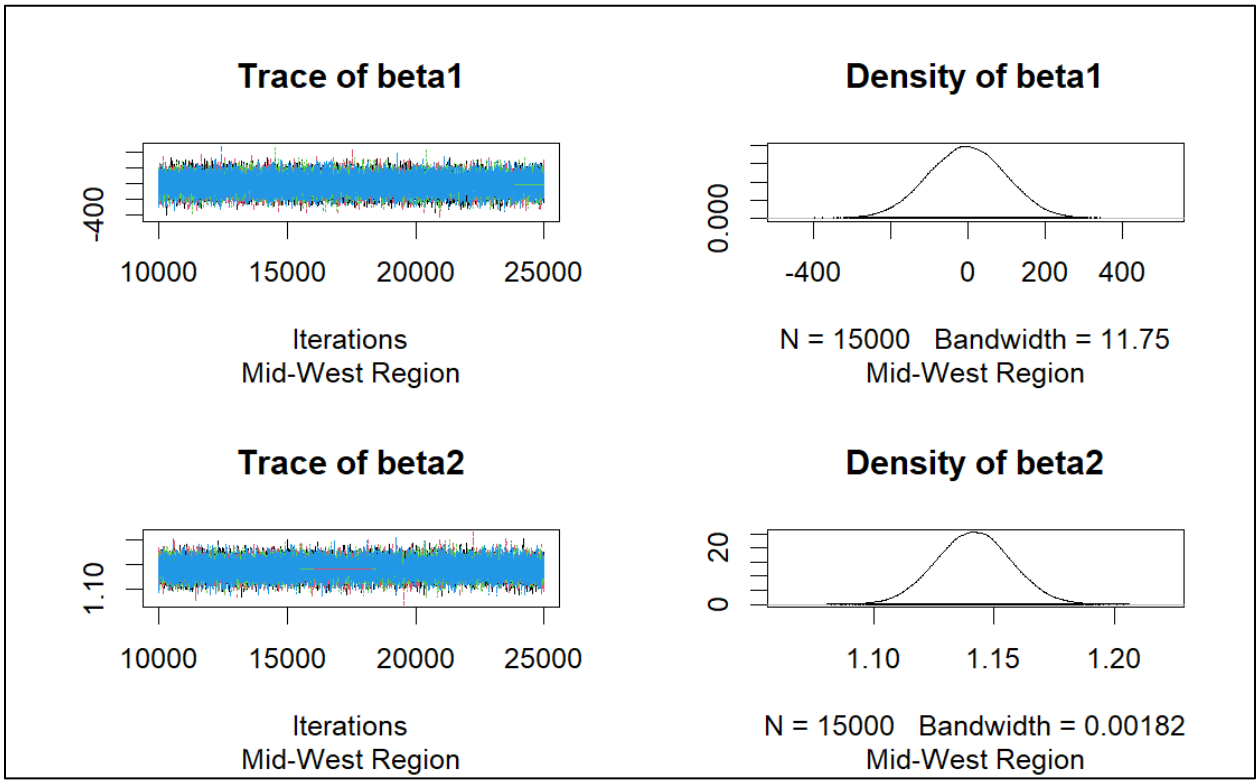

South East Region
Frequency
0
5
10
15
20
18000
20000
22000
24000
26000
Observed Count

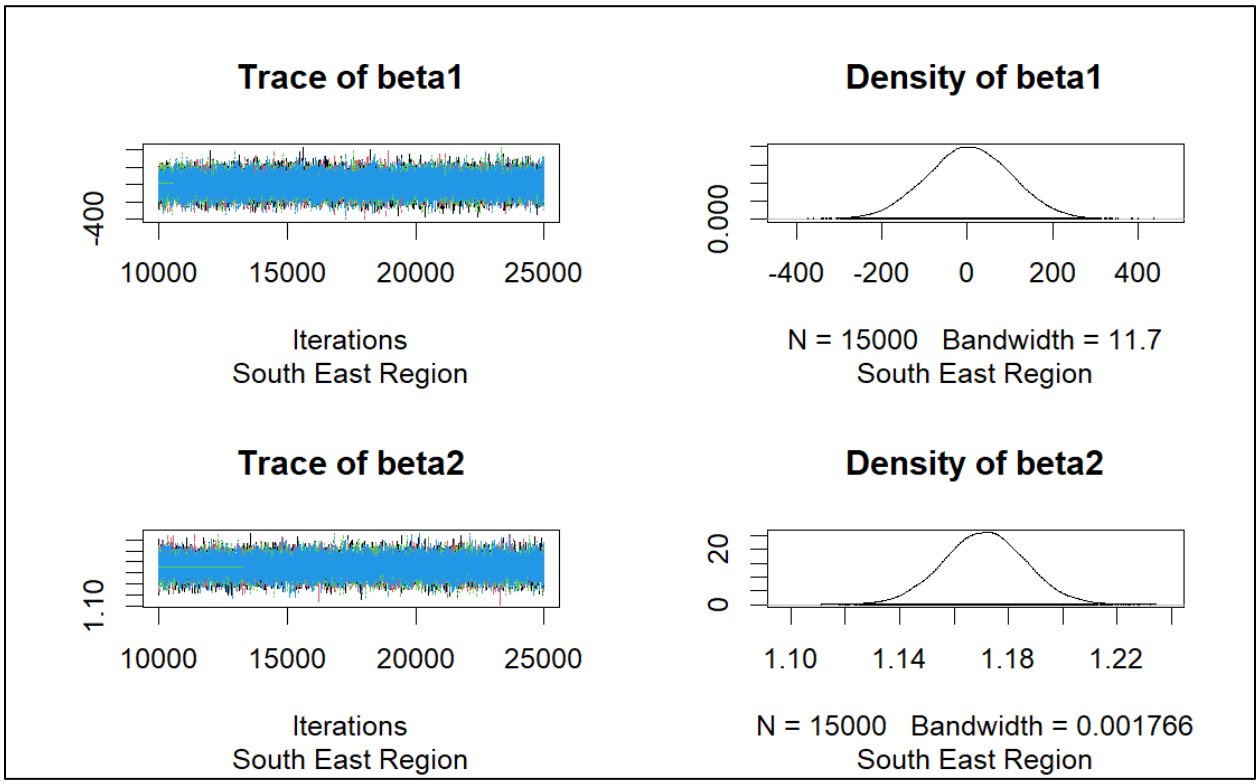

North East Region
Frequency
0
2
4
6
8
10
8200
8400
8600
8800
9000
9200
9400
9600
Average Expected Count

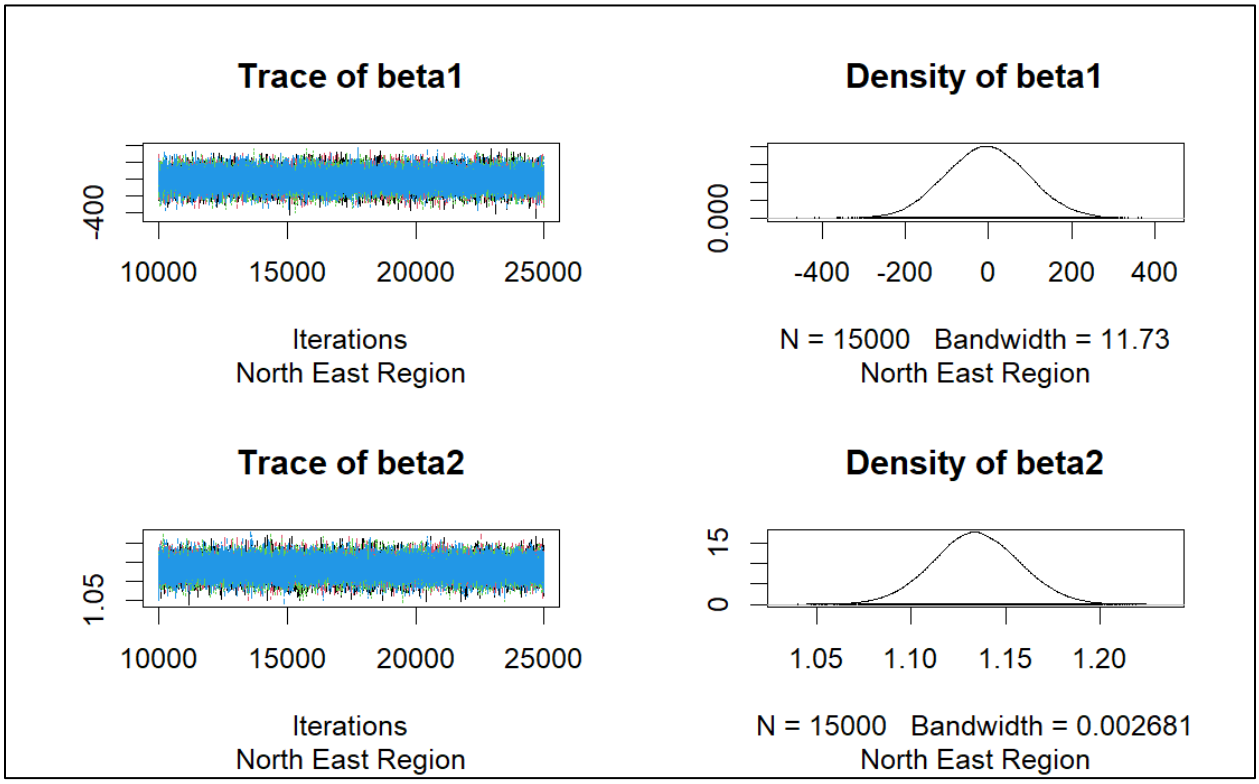

North East Region
Frequency
0
5
10
15
20
8000
10000
12000
14000
16000
18000
Observed Count


Trace of beta1
-20
10000
12000
14000
Iterations
West Region
Density of beta1
0.00
-40
-20
0
20
40
N = 5000 Bandwidth = 1.47
West Region
Trace of beta2
1.10
10000
12000
14000
Iterations
West Region
Density of beta2
0
20
1.10
1.14
1.18
1.22
N = 5000 Bandwidth = 0.002285
West Region


Trace of beta1
-400
10000
15000
20000
25000
Iterations
South West Region
Density of beta1
0.000
-400
-200
0
200
400
N = 15000 Bandwidth = 11.7
South West Region
Trace of beta2
1.15
10000
15000
20000
25000
Iterations
South West Region
Density of beta2
0
15
1.15
1.20
1.25
1.30
1.35
N = 15000 Bandwidth = 0.002886
South West Region


Trace of beta1
-400
10000
15000
20000
25000
Iterations
Mid-West Region
Density of beta1
0.000
-400
0
200
400
N = 15000 Bandwidth = 11.75
Mid-West Region
Trace of beta2
1.10
10000
15000
20000
25000
Iterations
Mid-West Region
Density of beta2
0
20
1.10
1.15
1.20
N = 15000 Bandwidth = 0.00182
Mid-West Region


Trace of beta1
-400
10000
15000
20000
25000
Iterations
South East Region
Density of beta1
0.000
-400
-200
0
200
400
N = 15000 Bandwidth = 11.7
South East Region
Trace of beta2
1.10
10000
15000
20000
25000
Iterations
South East Region
Density of beta2
0
20
1.10
1.14
1.18
1.22
N = 15000 Bandwidth = 0.001766
South East Region


Trace of beta1
-400
10000
15000
20000
25000
Iterations
North East Region
Density of beta1
0.000
-400
-200
0
200
400
N = 15000 Bandwidth = 11.73
North East Region
Trace of beta2
1.05
10000
15000
20000
25000
Iterations
North East Region
Density of beta2
0
15
1.05
1.10
1.15
1.20
N = 15000 Bandwidth = 0.002681
North East Region

West Region Model Summary

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

        Mean       SD  Naive SE Time-series SE
beta1 0.0207 10.05422 0.0710941      0.0722562
beta2 1.1627  0.01562 0.0001105      0.0001107

2. Quantiles for each variable:

         2.5%    25%      50%   75%  97.5%
beta1 -19.800 -6.760 -0.03946 6.819 19.653
beta2   1.132  1.152  1.16269 1.173  1.193
```

Mid-West Region Model Summary

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

       Mean       SD  Naive SE Time-series SE
beta1 6.143 99.63472 0.4067570      0.6234379
beta2 1.226  0.02459 0.0001004      0.0001539

2. Quantiles for each variable:

          2.5%    25%   50%    75%   97.5%
beta1 -189.062 -60.92 6.158 73.927 200.447
beta2    1.178   1.21 1.226  1.243   1.274
```

Southwest Region Model Summary

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

        Mean        SD  Naive SE Time-series SE
beta1 -2.229 100.06834 0.4085273      5.315e-01
beta2  1.142   0.01551 0.0000633      8.218e-05

2. Quantiles for each variable:

          2.5%     25%    50%    75%   97.5%
beta1 -199.126 -70.099 -1.944 65.530 192.113
beta2    1.111   1.131  1.142  1.152   1.172
```

Southeast Region Model Summary

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

       Mean       SD  Naive SE Time-series SE
beta1 4.596 99.61631 4.067e-01      4.666e-01
beta2 1.171  0.01515 6.183e-05      7.149e-05

2. Quantiles for each variable:

          2.5%     25%   50%    75%   97.5%
beta1 -190.541 -61.458 4.762 72.340 199.199
beta2    1.141   1.161 1.171  1.181   1.201
```

Northeast Region Model Summary

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

        Mean       SD  Naive SE Time-series SE
beta1 -4.151 99.90494 4.079e-01      0.5262079
beta2  1.134  0.02295 9.368e-05      0.0001201

2. Quantiles for each variable:

          2.5%     25%    50%   75%  97.5%
beta1 -197.879 -71.597 -4.536 63.18 193.43
beta2    1.089   1.119  1.134  1.15   1.18
```

Observed	Predicted
12603	12137.81
13496	12235.48
14311	12308.73
14646	12394.78
14585	12442.45
14265	12477.33
13701	12492.45
12470	12462.21
11836	12438.96
11539	12366.87
10985	12305.24
10650	12237.81
10409	12171.53
10019	12104.09
10330	12009.91
10366	11912.24
10233	11782.01
10315	11694.81
10203	11576.21
10168	11516.91
10322	11445.98
10498	11375.05
10381	11327.38
10539	11271.57
10348	11222.73
10564	11168.09
10586	11134.37
10680	11072.74
10562	11041.35
10853	11029.72
10568	11000.65
10811	10972.75
10586	10943.68
10672	10966.93
10227	10975.07
10408	10978.56
10862	10998.33
10418	11043.67
10479	11072.74
10298	11123.9
10157	11169.25
10172	11208.78
10273	11269.24
10405	11332.03
10813	11456.45
11015	11544.81
11112	11634.34
11650	11736.67
12244	11851.78
12528	12002.93
12481	12118.04
12667	12221.53
12430	12309.9

West

Observed	Predicted
18004	15492.17
19204	15549.24
19851	15617.73
19018	15675.95
18432	15711.33
17284	15721.6
16532	15705.62
15342	15674.8
14550	15655.4
13915	15612.02
13558	15541.25
13426	15424.82
12810	15296.98
12827	15188.54
12722	15058.42
12799	14953.4
12621	14832.41
12635	14731.96
12915	14624.66
12910	14527.64
12590	14445.45
12518	14367.83
12749	14283.36
12614	14228.57
12565	14154.38
12961	14116.71
12798	14073.34
12662	14044.8
12837	14028.82
12882	14013.98
12864	14016.26
13056	14034.53
12815	14060.78
12851	14068.77
12986	14081.33
12984	14104.16
12874	14145.25
12922	14217.16
12998	14286.79
13164	14358.7
13317	14448.88
13450	14549.32
13795	14630.37
13505	14746.8
13689	14830.12
13638	14916.87
14040	15013.9
14430	15134.89
14531	15246.76
15057	15338.07
15141	15433.96
15321	15542.39
15723	15632.57

Midwest

Observed	Predicted
8552	8471.186
9268	8547.205
9835	8585.215
10517	8619.546
10803	8657.556
10336	8672.269
9864	8666.139
8843	8624.451
8087	8599.928
7976	8575.406
7433	8548.431
7442	8512.874
7127	8476.091
6913	8407.428
6835	8343.67
6853	8259.068
6651	8186.727
6703	8132.778
6685	8048.176
6809	7997.905
6740	7962.348
6549	7932.921
6618	7885.102
6848	7860.58
7029	7833.606
6750	7804.179
6763	7784.561
7026	7749.004
7267	7745.325
7374	7733.064
7206	7730.612
6897	7717.125
7206	7715.898
6768	7706.089
6801	7712.22
6863	7730.612
6833	7740.421
6821	7771.074
6927	7778.43
6668	7812.762
6819	7849.545
6821	7890.007
6940	7946.408
6807	7996.679
6972	8044.498
7003	8124.195
7184	8196.536
7583	8289.721
8060	8360.836
7877	8435.629
7690	8512.874
7924	8588.893
8216	8655.103

Southwest

Observed	Predicted
21797	21823.92
23219	21921.12
25597	22024.18
26068	22074.53
26992	22115.52
26073	22127.23
24692	22108.5
22995	22096.78
21146	22033.54
20513	21969.13
19586	21874.27
19067	21786.44
18386	21664.64
18458	21536.99
18160	21398.8
18076	21254.76
17758	21134.13
17898	21008.82
17803	20860.09
17768	20717.21
18161	20563.8
17731	20419.75
18054	20312.01
18492	20230.03
18210	20144.54
18457	20060.22
18053	20027.43
18222	19978.24
18220	19927.89
18818	19886.9
18418	19823.66
18743	19799.06
18394	19782.67
18511	19786.18
18527	19808.43
18365	19833.03
18425	19877.53
18403	19946.62
18494	20014.55
18374	20094.18
18852	20205.44
18711	20309.67
18819	20417.41
18901	20579.02
19031	20705.5
19036	20895.22
18954	21082.6
19957	21296.92
19774	21517.08
20337	21715
19623	21873.1
20360	22065.16
21725	22236.15

Southeast

Observed	Predicted
11993	10698.7
12946	10759.95
13488	10804.18
13702	10833.66
12655	10854.08
11938	10858.62
11171	10858.62
10204	10849.54
9539	10822.32
9516	10804.18
9337	10757.68
8946	10697.57
8938	10610.24
8922	10500.23
8902	10401.55
8822	10307.42
8583	10201.94
8855	10122.55
8928	10062.44
9294	9977.384
8902	9902.53
8865	9860.567
8880	9776.64
8681	9729.005
8779	9659.822
8687	9614.457
9012	9601.981
8646	9597.444
8711	9579.298
8918	9558.883
8766	9544.14
8760	9550.944
8744	9549.81
8750	9565.688
8791	9566.822
8562	9606.518
8739	9642.81
8809	9685.908
8924	9730.14
9109	9778.908
9422	9850.359
9545	9920.676
9319	9985.323
9469	10074.92
9618	10156.58
9386	10239.37
9364	10321.03
9900	10385.68
10106	10487.75
10200	10573.94
10356	10643.13
10585	10721.38
10934	10812.12

Northeast