Nicholas A. Grokhowsky
SUR6905
November 28, 2018

An analysis of Crime Data Using SAR(INLA) for Bayesian Inference

According to professor Hochmair's manuscript, "Spatio-temporal Analysis of Land Survey Equipment Thefts in South-East Florida," there is a need for the analysis and interpretation of burglaries, larcenies, and robberies of surveyors' equipment in southeast Florida.[9] The Florida Surveying and Mapping Society (FSMS) maintains a database to aggregate these crimes. Professor Hochmair identified that there is an uneven distribution of these crimes throughout Florida, and it is in southeast Florida where the largest proportion of these crimes occur.[9] Multiple attempts have been made to analyze the spatial and temporal effects, as well as the interaction of space and time, on these crimes. The first attempt to model these crimes was a negative binomial model. However, it was advised that the use of a negative binomial model is unstable when the sample mean is less than four. Also, it was recommended that a simultaneous autoregressive function (SAR) be explored. The second attempt to model this data was completed using a log-normal Poisson distribution with CrimeStat's SAR method. The SAR in CrimeStat did not show results that were usable. This is likely because CrimeStat uses the asymptotic method, Markov Chain Monte Carlo (MCMC), which is inefficient while predicting from a sparse Gaussian Markov random field (GMRF).[1] The inefficiency is caused by the high computational cost of the exact measured values from MCMC. The recommended approach for sparse GMRF is the more recently developed Integrated Nested LaPlace Approximation (INLA) method because it uses approximations to reduce the computational cost.[1] A large concern for using an approximation method is the errors created by the approximations, however it has been shown that the MCMC error and INLA error are frequently the same.[2] Furthermore, there are similar studies that have used SAR(INLA) for Bayesian inference over a sparse GMRF. In the paper, Spatiotemporal Suicide Risk in Germany: A Longitudinal Study, the researchers setup hierarchical Bayesian Poisson models using SAR(INLA) to model over a sparse GMRF with accurate results.[3] For these reasons, my purpose for analyzing the FSMS data for a third time is to apply a SAR model better suited to modeling sparse GMRF's, and determine whether space and time had a significant effect upon the survey crimes distributed in southeast Florida.

The initial step was to query the data and build a database suitable for the R-INLA function in R. However, before any queries took place it was helpful to view the distribution of the data on a map. This was done using the leaflet() function and a base map from Open Street Map created by 'Stamen Design,' figure 1.[4] After visualizing the distribution of points within polygons it was clear that the attributes of each data set needed to be joined. The join was done using the methods over(), mutate(), left_join(), group_by(), and arrange(). Next, the data frame was built and renamed, the crimes were separated by year, and the expected values were calculated. The expected values were calculated for each year by the product of the population and the sum of crimes per year and divided by sum of the populations. This queried data set left us with the attribute data for the crime points and tract polygons with expected crime values, and in wide format. However, the R-INLA function requires the data to be in long format. In order to facilitate this format change the reshape() function was used. The final result was a data frame with attribute values from crime points, tract polygons, and expected values in long format. The result was 1,360 rows of data with 37 variables. After the format change each year's expected values were merged into a single column, and the individual yearly expected value columns were removed. Finally, the densities were calculated for area, population, police stations, jobs sums, black population, homeowners, and renters. The data frame had 1,360 rows with 21 variables.

Nicholas A. Grokhowsky
SUR6905
November 28, 2018

In order to identify covariance and eliminate confounding variables a correlation matrix was created using the ‘Spearman’ method. The ‘Spearman’ method was used because the ‘Pearson’ method was designed for linear relationships, and is subject to false positives when comparing non-linear relationships.[8] Therefore, the ‘Spearman’ method was used because it accurately identifies monotonic relationships.[5] Also, in order to facilitate an easy to interpret visualization of the correlation matrix, a flattenCorrMatrix() was used from STHDA.[6] The flattenCorMatrix() method made it possible to plot the correlation matrix as a series of points that represent negative correlation in red and positive correlation in blue. Additionally, the points varied in size at the set, positive and negative, thresholds of 0, 0.3, 0.6, 0.8, 0.95 , and 1. The original correlation matrix can be seen in figure 2. After reviewing the correlation matrix the highly correlated variables were removed. The final correlation matrix can be seen in figure 3. It is important to note that there was a 0.46 correlation between black population density and renter density. Although this is not a strong correlation it is the highest correlation remaining, and it is something that was taken into consideration during the modeling of these crimes.

After eliminating the confounding variables from the data frame the ecological regression model was calculated. The response variable, Y, was made equal to the total crimes per year, and the remaining variables were set to identifiable variable names that corresponded to the final correlation matrix (figure 2). The first model was the null model: Y ~ 1. The DIC for the null model is 1508.21. Forward stepwise regression analysis identified a model with the random variables tracts, years, and income, and the fixed variables police station density and "treeness". The DIC was reduced to 1294.50. However, reanalyzing the model identified a last model where the year variable was removed from the model, and it further reduced the DIC to 804.34. Considering that the year variable was removed it was necessary to re-query the data frame again. The tract data (polygons) were simply merged with the crimes per year and no change to the data shape was necessary. The new data frame was modeled with a null model which has a DIC of 546.60, and the fitted model model has a DIC of 333.12 (figure 4).

Another measure for goodness of fit for the INLA method is the measure of Log Pseudo Marginal Likelihood (MPML) and the Probability Integral Transformation (PIT)[7]. The MPML can be calculated by the sum of the logged Conditional Predictive Ordinate (CPO), and larger values identify better fitted data[7]. The current model was applied to the data that excluded the temporal effect, and has a larger MPML of -265.369 compared to the model applied to the data set that includes the temporal effect of -651.25. Also, the histogram of the PIT showed a distribution that was closer to normal for the data set that excluded the temporal data compared with the histogram of the PIT for the data set that included the temporal data. For these reasons the final model was applied on the data set that excluded temporal data.

The most notable outcome of both models is the effect of 'treeness' on crimes. There is a moderately negative correlation between 'treeness' and crimes at -0.46 when the time data is removed from the data set, and -0.35 when the time data is left in the data set but removed from the model (figure 4 & 5). The significance of the 'treeness' effect is very high ($p < 0.001$) for the data set excluding the time data, and for the data set that includes time data but excludes it from the model (figure 4 & 5). 'treeness' was measured in the study as a ratio between 0 and 1, where the greater the measure the greater the road network branches. Alternately, the lower measurements reflect road networks that are closest to circuits. These models support one of professor Hochmair’s theories that locations along road networks that are closer to a circuit are more likely to experience these crimes.

Nicholas A. Grokhowsky
SUR6905
November 28, 2018

There is strong evidence to show that time did not have an effect on the crimes within the FSMS database. The initial modeling phase emphasized backwards stepwise regression analysis in order to include time, space, and the space-time interaction. These models consistently showed high DIC values. Furthermore, these models that did include time, space, and the time-space interaction showed no relationships between time and the fitted values. When switching approaches to forward stepwise regression analysis the time variable was removed, and it drastically reduced the DIC to 804.34. The final model used ‘tracts’ and ‘income’ as random effects, and 'treeness' and ‘police densities’ as fixed effects. The outcome identified a highly significant effect, with a moderate strength, coming from 'treeness', and no effect coming from time. The time trend is provided for the model that was applied to the data set which includes the temporal data, and it shows that there is no effect on these crimes based on when they occurred (figure 10). For these reasons, and the goodness of fit measures, it was a better option to use a data set that excludes the time variable.

When reviewing the output of the final model on the data set that excludes temporal data, it is clear that many of the tracts have a significantly higher probability for survey equipment crimes. The fitting results displayed in figure 9 shows that many of the tracts survey equipment crime rates are well above the confidence interval. A iteration over the data frame that compared observed values to predicted values above and below the confidence intervals was performed in order to identify tracts with significant crimes in either direction. These tracts are listed in the file TractID.dat. Additionally, maps of tracts with crimes greater than the 97.5% confidence interval and crimes less than the 2.5% confidence interval are provided (figure 9), along with maps of the observed, predicted, and residual values (figure 6 & 7). These maps show the locations of highest probability to be throughout the study area, and the locations of the lowest probability are more dispersed throughout the study area. It is important to clarify that the significance of these tracts is primarily caused by the 'treeness' effect.

Finally, although the model did not show a temporal relationship with the FSMS crimes, it did bring attention to the concept of 'treeness', which is a spatial measurement that was crafted by professor Hochmair. According to professor Hochmair’s manuscript, there have been studies that identified an increase in crime activity with networks of higher connected roadways. The use of 'treeness' as a measurement tool of connected roadway networks enabled the identification of an effect that was significant. The identification of the significance of this measurement creates other questions about roadway networks, and how their 'treeness' might effect other studies. In fact, this measure might be an important considerations when analyzing spatial data in other fields. It would be interesting to see the use of 'treeness' in the analysis of police routes, city transportation planning, emergency response, real estate, and retail site selection. After concluding this analysis it is clear that the time variable did not have an effect on the FSMS crimes, but the spatial effect of 'treeness' did. This outcome shows that there is a need to continue research into the spatial measurement theory of 'treeness'.

Nicholas A. Grokhowsky
SUR6905
November 28, 2018

Figure 1: Visualization of crime points and tract polygons on base map of Florida

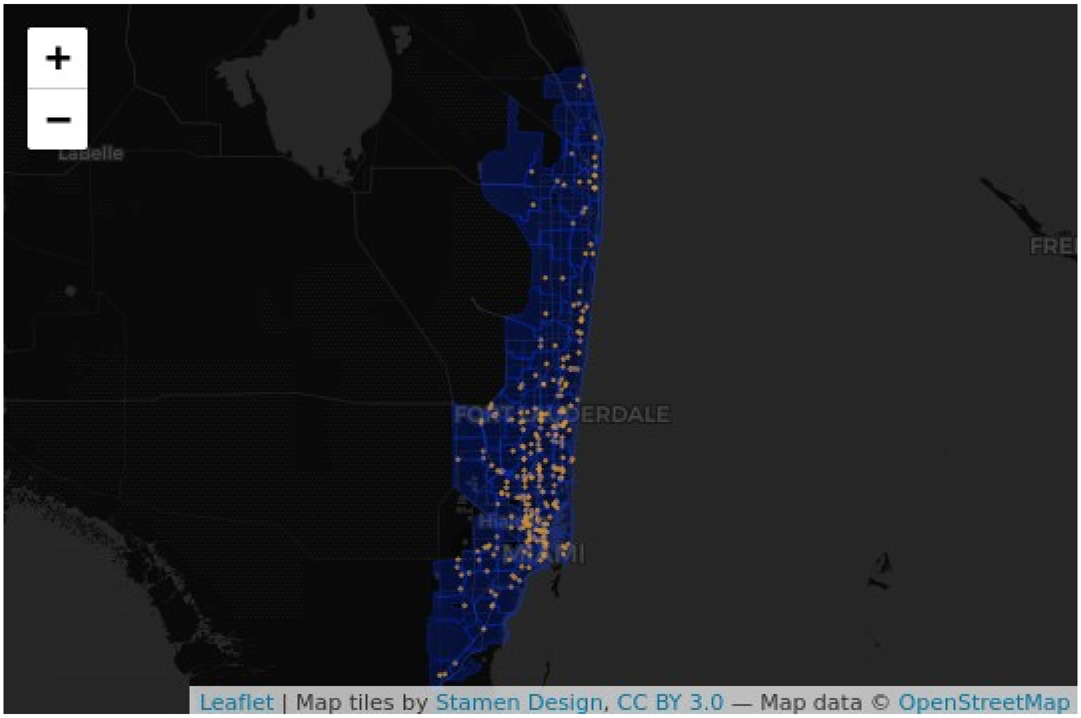
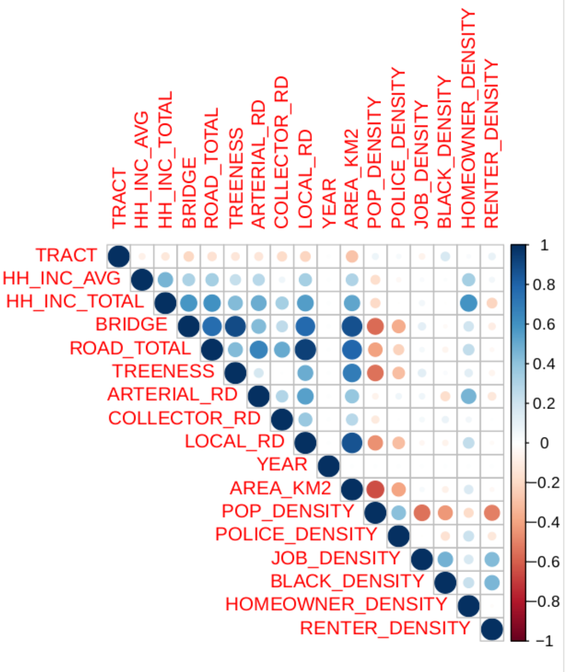

+
−
Labelle
FRE
FORT LAUDERDALE
MIAMI
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap

Figure 2: Original correlation matrix visualization used to identify confounding variables



Nicholas A. Grokhowsky
SUR6905
November 28, 2018

Figure 3: Final correlation matrix after confounding variables were removed

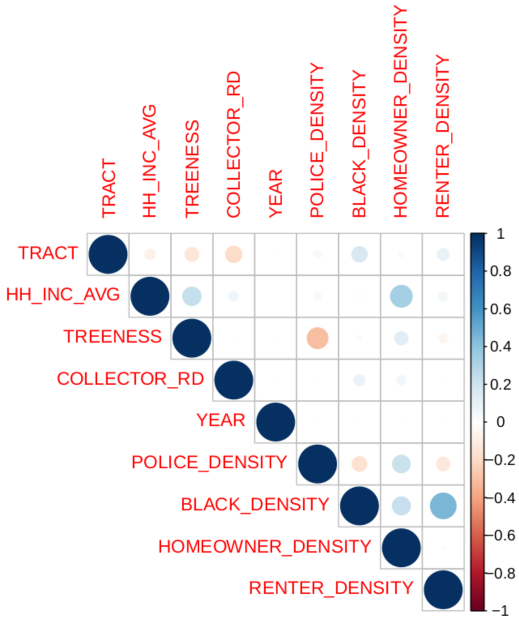
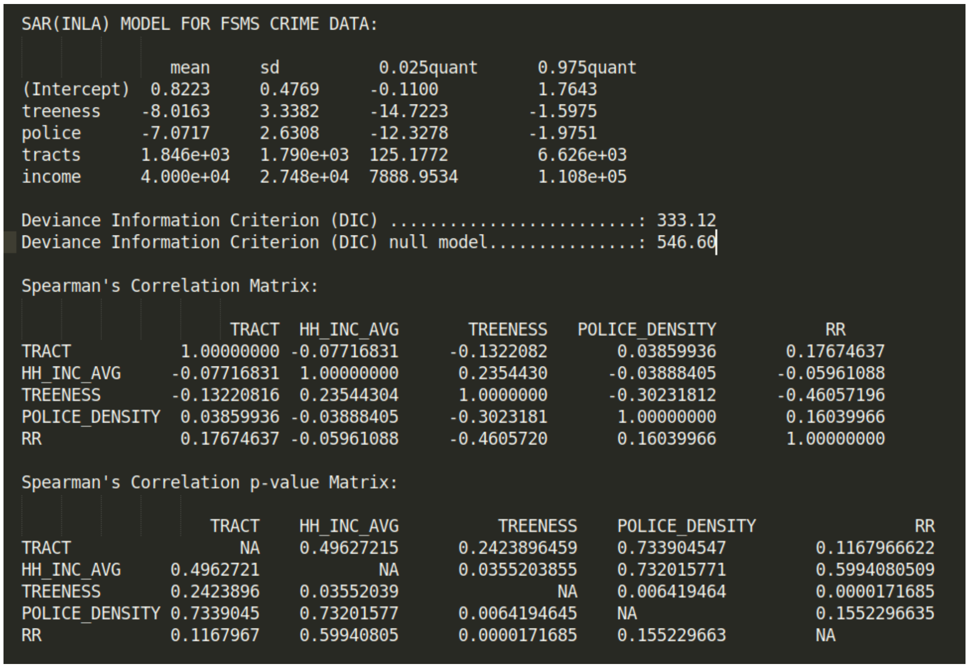


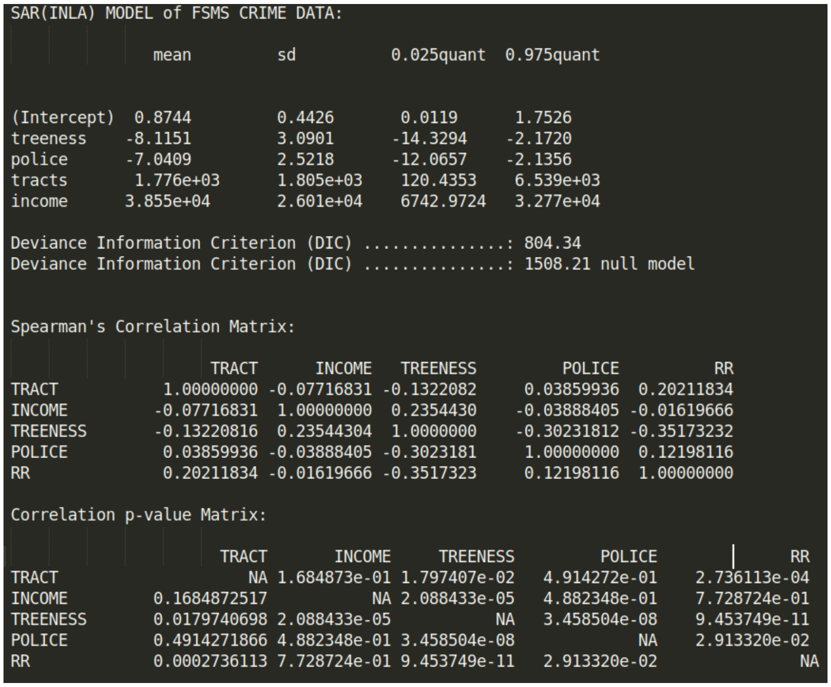
Figure 4: Final SAR(INLA) model on on data set that includes temporal data



```
SAR(INLA) MODEL FOR FSMS CRIME DATA:

               mean     sd          0.025quant      0.975quant
(Intercept)  0.8223     0.4769     -0.1100          1.7643
treeness    -8.0163     3.3382     -14.7223        -1.5975
police      -7.0717     2.6308     -12.3278        -1.9751
tracts      1.846e+03   1.790e+03  125.1772         6.626e+03
income      4.000e+04   2.748e+04  7888.9534        1.108e+05

Deviance Information Criterion (DIC) .........................: 333.12
Deviance Information Criterion (DIC) null model...............: 546.60

Spearman's Correlation Matrix:

                     TRACT  HH_INC_AVG       TREENESS   POLICE_DENSITY           RR
TRACT           1.00000000 -0.07716831     -0.1322082       0.03859936       0.17674637
HH_INC_AVG     -0.07716831  1.00000000      0.2354430      -0.03888405      -0.05961088
TREENESS       -0.13220816  0.23544304      1.0000000      -0.30231812      -0.46057196
POLICE_DENSITY  0.03859936 -0.03888405     -0.3023181       1.00000000       0.16039966
RR              0.17674637 -0.05961088     -0.4605720       0.16039966       1.00000000

Spearman's Correlation p-value Matrix:

                   TRACT    HH_INC_AVG         TREENESS    POLICE_DENSITY                RR
TRACT                 NA    0.49627215      0.2423896459    0.733904547         0.1167966622
HH_INC_AVG     0.4962721            NA      0.0355203855    0.732015771         0.5994080509
TREENESS       0.2423896    0.03552039                NA    0.006419464         0.0000171685
POLICE_DENSITY 0.7339045    0.73201577      0.0064194645    NA                  0.1552296635
RR             0.1167967    0.59940805      0.0000171685    0.155229663         NA
```

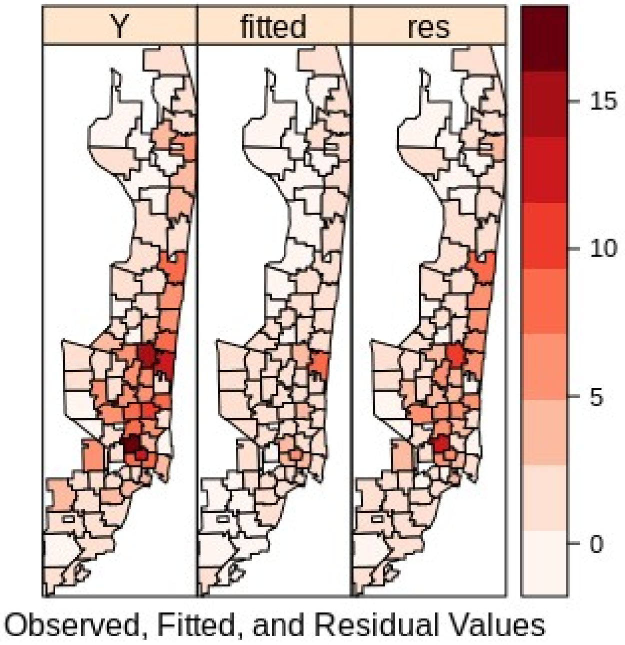
Nicholas A. Grokhowsky
SUR6905
November 28, 2018

Figure 5: Final SAR(INLA) model on data that excludes temporal data



```
SAR(INLA) MODEL of FSMS CRIME DATA:

              mean         sd          0.025quant  0.975quant


(Intercept)  0.8744         0.4426       0.0119      1.7526
treeness    -8.1151         3.0901      -14.3294    -2.1720
police      -7.0409         2.5218      -12.0657    -2.1356
tracts       1.776e+03      1.805e+03    120.4353    6.539e+03
income      3.855e+04       2.601e+04    6742.9724   3.277e+04

Deviance Information Criterion (DIC) ...............: 804.34
Deviance Information Criterion (DIC) ...............: 1508.21 null model


Spearman's Correlation Matrix:

                     TRACT      INCOME   TREENESS         POLICE          RR
TRACT           1.00000000 -0.07716831 -0.1322082     0.03859936  0.20211834
INCOME         -0.07716831  1.00000000  0.2354430    -0.03888405 -0.01619666
TREENESS       -0.13220816  0.23544304  1.0000000    -0.30231812 -0.35173232
POLICE          0.03859936 -0.03888405 -0.3023181     1.00000000  0.12198116
RR              0.20211834 -0.01619666 -0.3517323     0.12198116  1.00000000

Correlation p-value Matrix:

                      TRACT       INCOME     TREENESS         POLICE          RR
TRACT                    NA 1.684873e-01 1.797407e-02   4.914272e-01    2.736113e-04
INCOME         0.1684872517           NA 2.088433e-05   4.882348e-01    7.728724e-01
TREENESS       0.0179740698 2.088433e-05           NA   3.458504e-08    9.453749e-11
POLICE         0.4914271866 4.882348e-01 3.458504e-08             NA    2.913320e-02
RR             0.0002736113 7.728724e-01 9.453749e-11   2.913320e-02              NA
```

Figure 6: Observed crimes compared to predicted crimes (excludes temporal data)



Nicholas A. Grokhowsky
SUR6905
November 28, 2018

Figure 7: Posterior mean spatial effects

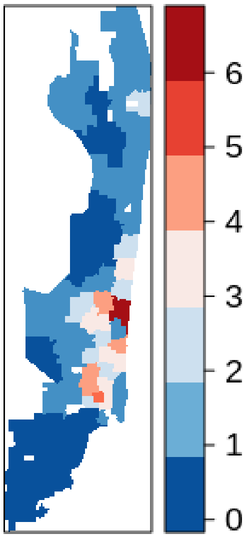
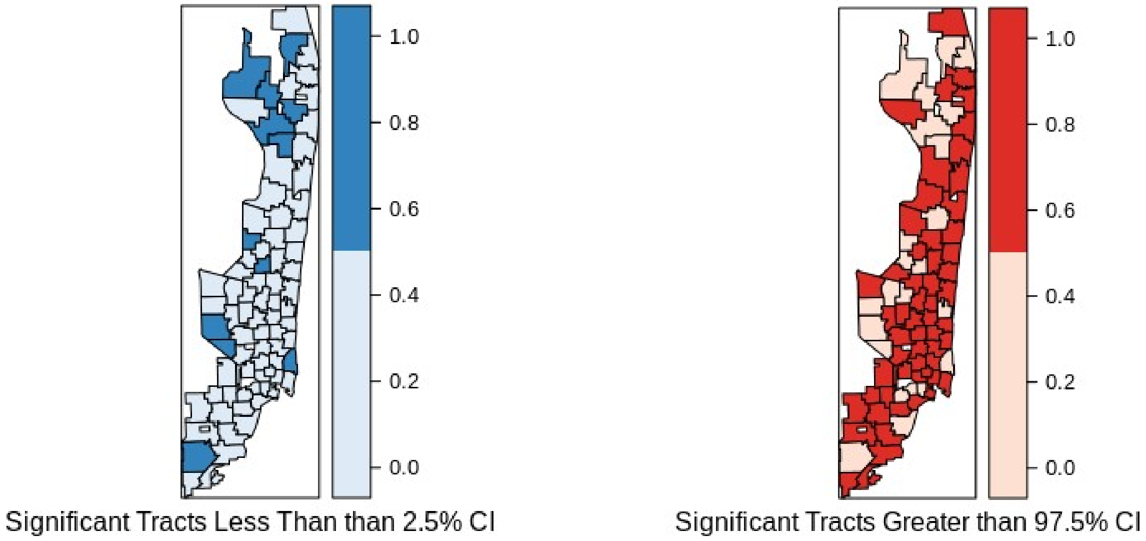


Figure 8: Maps of significant crime values based on 2.5% and 97.5% quantiles (excludes temporal data)


Significant Tracts Less Than than 2.5% CI

Significant Tracts Greater than 97.5% CI

Nicholas A. Grokhowsky
SUR6905
November 28, 2018

Figure 9: Fitting results for data that excludes temporal data

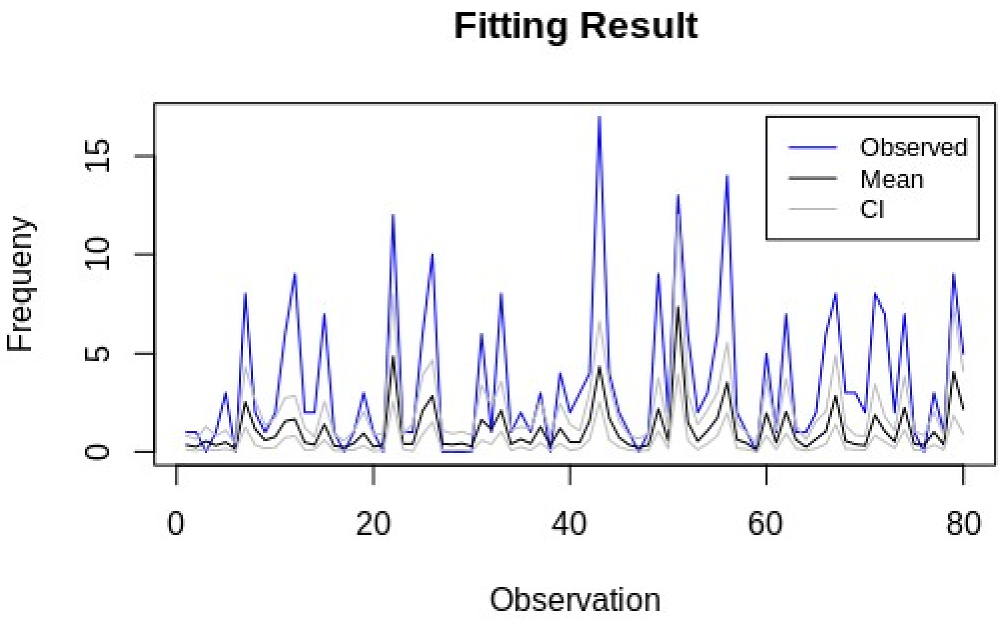
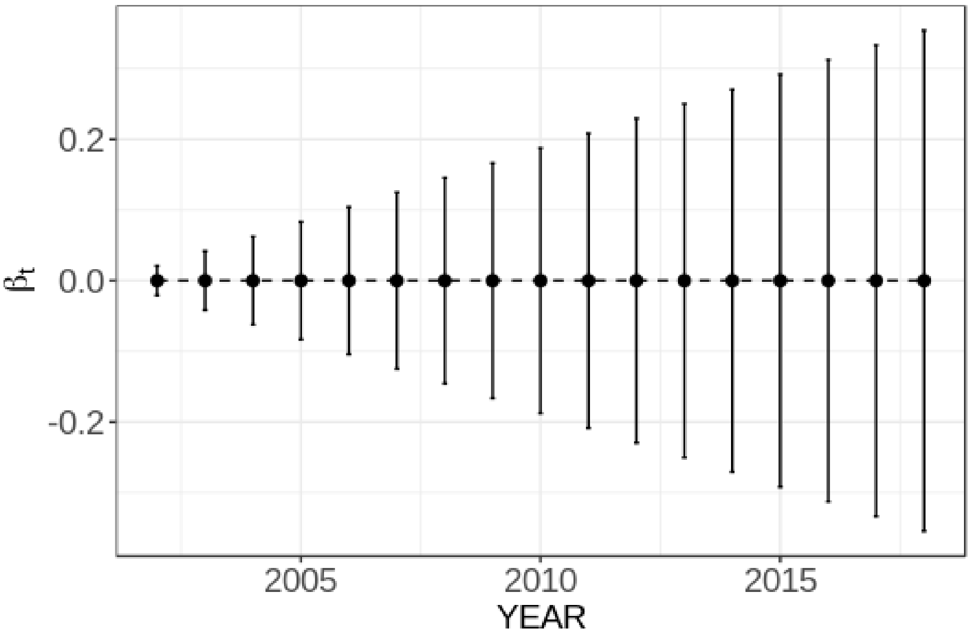


Figure 10: Time trend for model of data which includes temporal data



Nicholas A. Grokhowsky
SUR6905
November 28, 2018

1. Taylor, Benjamin M and Peter J Diggle, “INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in log-Gaussian Cox Processes,” March 20, 2012

2. Morrison, Kathryn, “A Gentle INLA Tutorial,” https://www.precision-analytics.ca/blog-1/inla, December 20, 2017

3. Helbich, Marco, Paul L Plener, Sebastian Hartung, & Victor Blumi, “Spatiotemporal Suicide Risk in Germany: A Longitudinal Study, August 9, 2017

4. Rodenbeck, Eric, Stamen, https://stamen.com/

5. MiniTab, Inc, “A Comparison of Pearson and Spearman Correlation Methods,” https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/, 2017

6. STHDA, “Correlation matrix : A quick start guide to analyze, format and visualize a correlation matrix using R software,” http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

7. Xiaofeng Wang, Yu Yue Ryan, Julian J. Faraway, Bayesian Regression Modeling with INLA CRC Press, February 16, 2018

8. Pernet, Cyril, Rand Wilcox, and Guillaume A. Rousselet, “Robust Correlation Analysis: false positive and poer validation using a new open source Matlab toolbox,” https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00606/full, January 10, 2013

9. Hochmair, Hartwig, “Spatio-temporal Analysis of Land Survey Equipment Thefts in South-East Florida,” FOR PEER REVIEW ONLY