Spatial Pattern of Population Movement Using Cellular Data

Wenyu Gao & Danni Lu
Department of Statistics, Virginia Tech

Abstract

As increasing number of people are living in urban area, rising issues such as deteriorating periodical traffic congestion in central business districts is gaining more and more attention. The fundamental question to alleviate traffic congestion in urban area is to figure out where the population origins and how population moves spatially. In this report, we applied a spatiotemporal model to analyze the characteristics of population density in metropolitan based on cellular data. Specifically, taking Shanghai as a case study, the spatial and temporal pattern of population density was discussed. Different spatial models were fitted and compared. Predictions were made based on selected models and the results are displayed in choropleth map. The population movement pattern is discussed based on the map. Results show that during the morning peak hours, population moves from all over the city to central areas. Areas with relative high population density expand, grow and converge along transportation corridors.

Key words: Cellular Data, Likelihood, Periodical Traffic, Population Movement, Semivariogram, Spatial Analysis.

1 Introduction

Worldwide, urban population has been growing steadily during the past decades, from 33.56% of the world's population in 1960 to 53.86% in 2015. In developed countries, the proportion is even higher. 82% of total population in United States was from urban areas in 2015(World Bank, 2015). As increasing number of people are living in urban areas, the rising issues such as limited land resource and deteriorating recurrent traffic congestion are repeatedly emphasized by transportation planning institutes. One of the major causes of recurrent traffic congestion is high intensity development in districts where the traffic attraction exceeds the accommodation volume of traffic system. Central Business District (CBD) is therefore becoming the key area to congestion solution as it has been bringing remarkable pressure to transportation system during peak hours on weekdays due to their high traffic attraction (Willett K, 2006).

Central Business District (CBD) is a major destination for urban area. According to a traffic survey of 63 cities in US in the 1950s, about one in every five metropolitan residents has at least one destination in the CBD during each weekday (Foley D L,

1952). Integration of multiple services within a small area is convenient not only for people living in the vicinity but also for people living elsewhere in the city. However, the price to pay is that area with higher developing density requires transportation system with larger capacity during peak hours (Mindali O, 2004), especially in big metropolitan like New York, Sydney and Shanghai. As a consequence, traffic congestion in CBD areas has become a bottleneck for every growing metropolitan to serve better city life.

Researchers have been seeking for solutions to alleviating recurrent traffic congestion for years. Sufficient work has been done on analyzing spatial and temporal characteristic of travel behavior during peak hours. Traditional method to acquire travel pattern information is to conduct household survey, which is usually inefficient and expensive. Nowadays, high density of cellular towers and deep penetration of wireless service provides us with detailed travel information with high accuracy in precise timings. Previous studies have shown a promising future of mobile data in exploring commuting traffic. Compared with traditional way of exploring travel pattern, cellular data stands out for wider coverage area, better representative of population with all kinds of travel pattern, and more detailed information with high accuracy and precise timing. Thus, cellular data allows us to study temporal and spatial travel characteristics of higher graininess.

Midtown in Shanghai is one of the largest central business districts in the world, which makes it a suitable place to study travel pattern related to CBD. Based on cellular data collected in Shanghai, we're going to discuss spatial and temporal characteristics of travel pattern related to CBD in Shanghai. In this study, we treat the study area as a continuous field. In section 2, we give a brief introduction of our data, the goal of our study, and make some exploratory analysis to prepare spatial data for modeling. In section 3, we introduce the methodology adopted to build, evaluate and apply the spatial models we fit. Afterwards, we demonstrate and discuss the results of our empirical study, model fitting and prediction in section 4. At last, conclusions are drawn from the results and discussion in section 5.

2 Data Description

We collected geodetic position information of all users in Shanghai from China Mobile for a continuous 24-hour period. China Mobile is the biggest carrier in China, with 64% total ownership in China (China Mobile, China Unicom and China Telecom, 2016). The dataset includes coded mobile phone ID, date, time, latitude and longitude of corresponding cellular tower. Cellular towers are infrastructure that enables voice and data services for mobile phones. They operate in different radio frequencies, and allow users to maintain their connections while traveling from one base station to another. In total, we collected more than 1.1 billion cell tower hands-off records. These records are from 37,450 different cell towers all over Shanghai with average density of 5.91 cell tower per kilometer square.

One issue of adopting cellular data is privacy of participants (Steenbruggen, John, 2013). To protect their privacy, all records are anonymized and replaced by an identifier ID before we got access to it. In addition, records were aggregated by areal unit and time period before analyzing and visualizing. No personal information is involved and displayed in the research.

The spatial pattern of population density has a direct relationship with traffic demand. Every morning on weekday, metropolitan sees a population movement from every corner of the city to CBD areas. During peak hours, the traffic on roads and crowding transport is at its highest. The huge increment of population brings tremendous pressure to transportation system near CBD areas. In traditional traffic survey, no standard areal definition of CBD was followed (Foley D L., 1952). Though the importance of CBD in transportation is self-evident, it is much more of a general concept in transportation study. Luckily, the comprehensive information of cellular data provides us a way to identify CBD by its transportation characteristics. In this study, we focus on main districts that are circled by Outer Expressway in Shanghai (Figure 1).

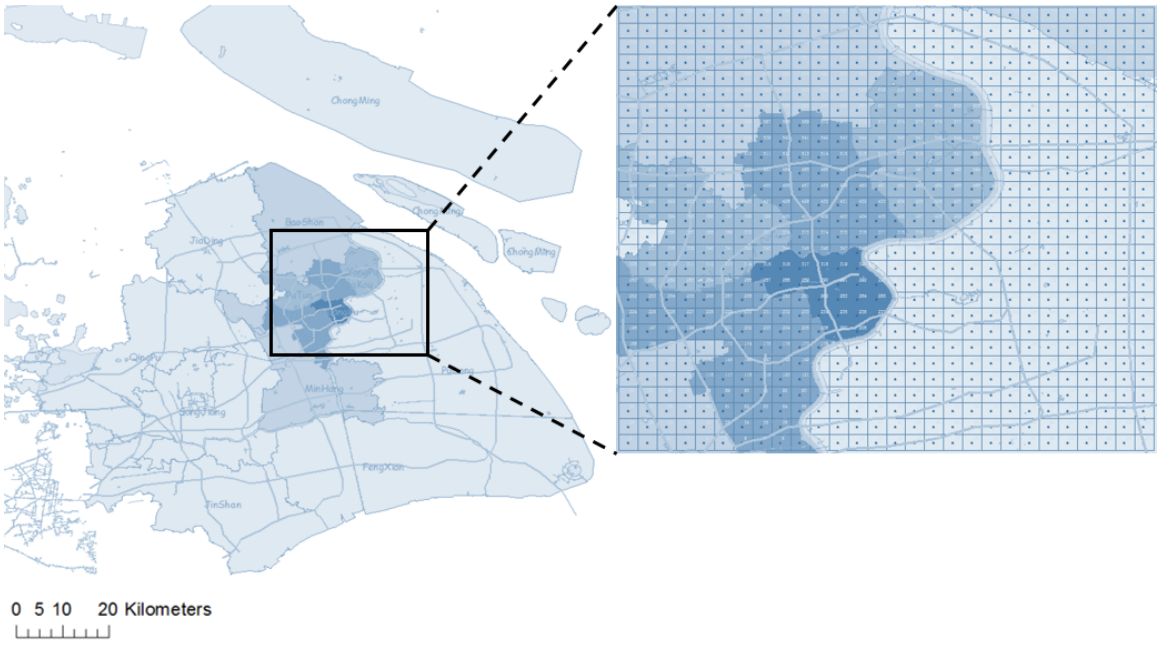

Figure 1: Study Area and Aggregation Grid

The accuracy of location information we get depends on the density of cell towers. Instead of accurate location of mobile phone users, cellular data records location of cell tower that is nearest to the mobile phone users. As we can see in Figure 2, cell towers are not evenly spaced. The density in the center area is higher than outer areas. As a result, the coverage area of each cellular tower is different. To simplify the situation, study area is partitioned into a 28*28 grid, each areal unit is one square kilometer. All cellular information from one areal unit is aggregated into one fictional cell tower in the geo-center of the areal unit (Figure 2). Hence, we have a regular point-referenced data to establish spatial model. In terms of study period, our goal is exploring spatial and temporal pattern of population in rush hours in the morning. Generally, typical morning rush hours for metropolitan CBDs are between 7:00 and 9:00 am. Therefore, we extract cellular data from 5:00 am to 10:00 am to cover this range.

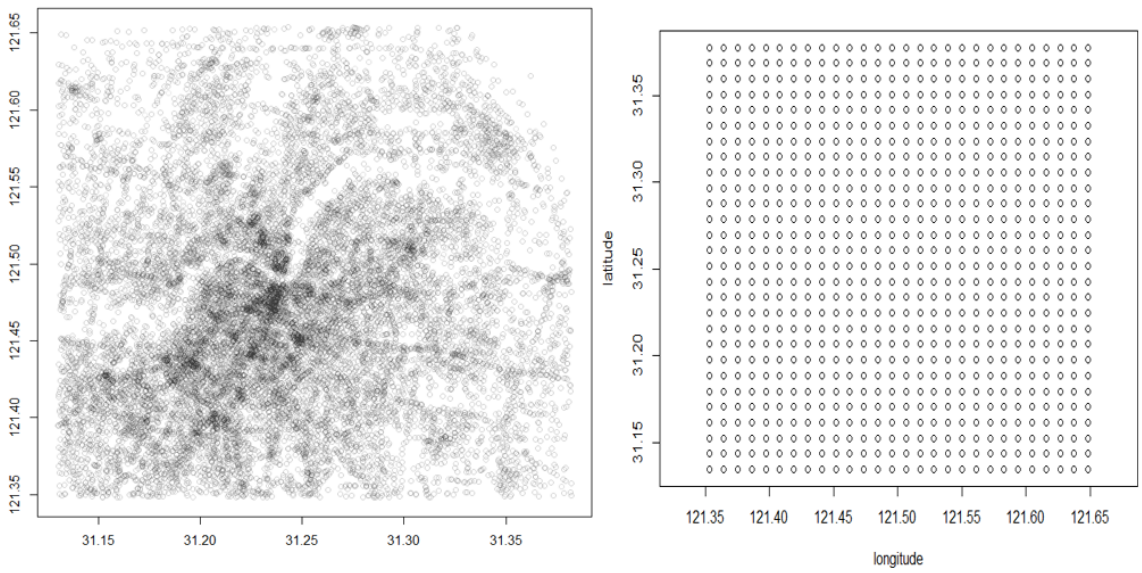

Figure 2: Original Cell Towers and Aggregated Cell Towers

3 Methodologies

Based on our data characteristics and aggregation methods, we treat our data as point-referenced data and apply the corresponding spatial theories for analysis. The whole analyzing process is as follows: We begin from exploratory analysis to check if there is any special structure of the data, followed by identification of spatial trend. We would like to analyze the data without spatial trends because this will reflect the intrinsic covariance structure of the data. After detrending, we will check for existence of anisotropy. Model fitting will be conducted after all these preliminary work has been done. Semivariogram fitting and likelihood fitting models are considered. Some model selection criteria are applied to select the best fit and predicted model. Last but not least, we will make inference on the predicted models.

Most of our studies relate only to population density with locations and time. We only consider the two covariates: distance to Metro and distance to Expressway in model fitting and is also separate from models with population density. We study the effects of the covariates by fitted models.

3.1 Preliminary Analysis

All studies in preparation for model fitting are owed to preliminary analyses. Most of the preliminary checks are based on visualization, except for spatial trend detection, where model fitting is conducted. The trend is mainly diagnosed with coordinates. Stepwise selection starting with 2nd order model is used to select the best model for each hour. AIC is used as criterion.

In the preliminaries, we check both frequency variable and logarithm of the variable. For trend analysis, we also plot the boxcox transformation, which also suggesting a logarithm transformation. However, as shown in the results, taking logarithm will not help improve the result greatly. We tend to use the original data.

3.2 Model Fitting

When plotting the semivariogram, the plot is not reliable if the distance is too large. In data preparation, we aggregated the data to a square of 28km by 28km, indicating that the largest distance is the diagonal line. However, as people are more concentrated in the center, the distance of interest should not be too large. Thus in the following model fitting, we select the largest distance to be around half of the edges, i.e 14 km.

3.3 Prediction

Model selection for best predicted model is conducted by both visualization and residual sum of squares comparison. Predicted images are based on predicted values at the interpolated points when plotting image for true data rather than the recorded places of true data in the study area. Thus we can compare the predicted models to the true data.

4 Results

4.1 Exploratory Data Analysis

We begin with a summary table of the dataset, shown in Table 1. From the table, it

Table 1: Summary Table of Dataset

	FID	long	lat	freq
1	Min. : 0.0	Min. :121.4	Min. :31.13	Min. : 0
2	1st Qu.:195.8	1st Qu.:121.4	1st Qu.:31.20	1st Qu.: 2060
3	Median :391.5	Median :121.5	Median :31.26	Median : 5894
4	Mean :391.5	Mean :121.5	Mean :31.26	Mean : 9773
5	3rd Qu.:587.2	3rd Qu.:121.6	3rd Qu.:31.32	3rd Qu.:14575
6	Max. :783.0	Max. :121.6	Max. :31.38	Max. :83822
	D2Metro	D2Road	hour	
1	Min. : 0.395	Min. : 2.274	Min. :5	
2	1st Qu.: 387.974	1st Qu.: 397.764	1st Qu.:6	
3	Median : 973.144	Median : 888.332	Median :7	
4	Mean :1373.922	Mean :1126.451	Mean :7	
5	3rd Qu.:1981.959	3rd Qu.:1609.672	3rd Qu.:8	
6	Max. :6852.305	Max. :5607.713	Max. :9	

is clear to see that the locations do not vary greatly in longitude and latitude. The frequency, distance to Metro and distance to road have large range and do not seem to be symmetric. There seems no problematic data. We then focus on the frequency variable and we will check the image plot (Figure 3), perspective plot (Figure 21 in Appendix A), histogram (Figure 4), boxplot (Figure 22 in Appendix A) and qqplot (Figure 24 in Appendix A) on the variable. Due to page limitation, we only present the image plot and histograms in the paper. The other plots are put to appendices.

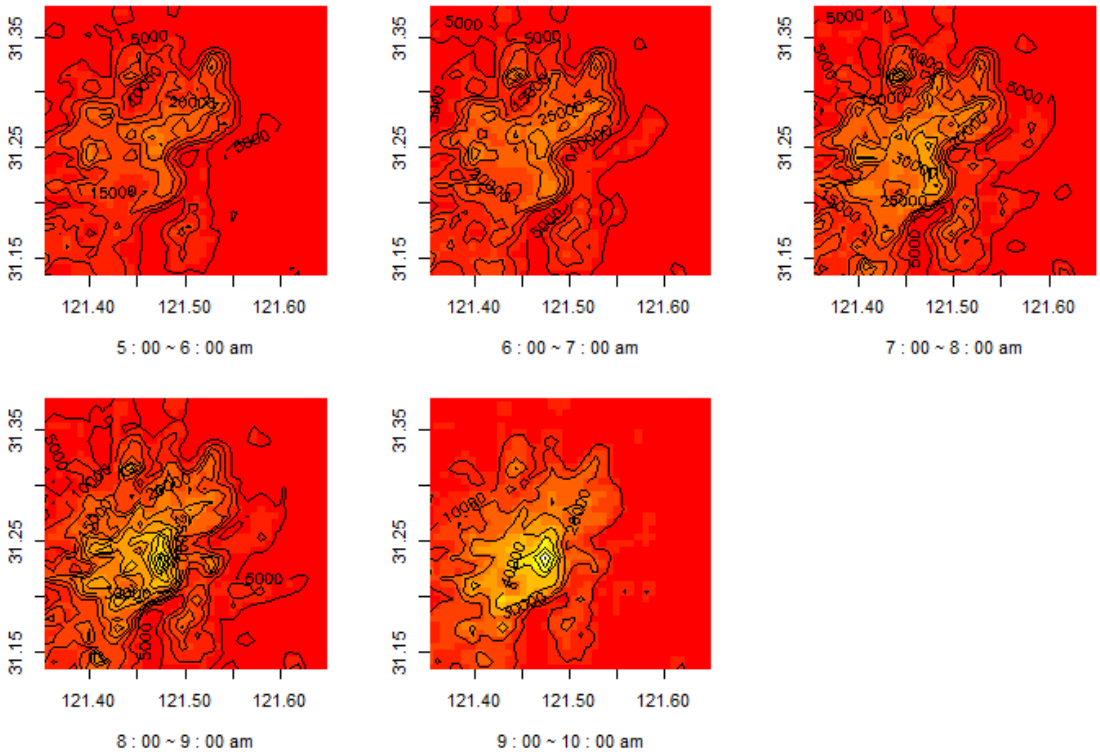

Figure 3: Image Plots of Frequency by Hour

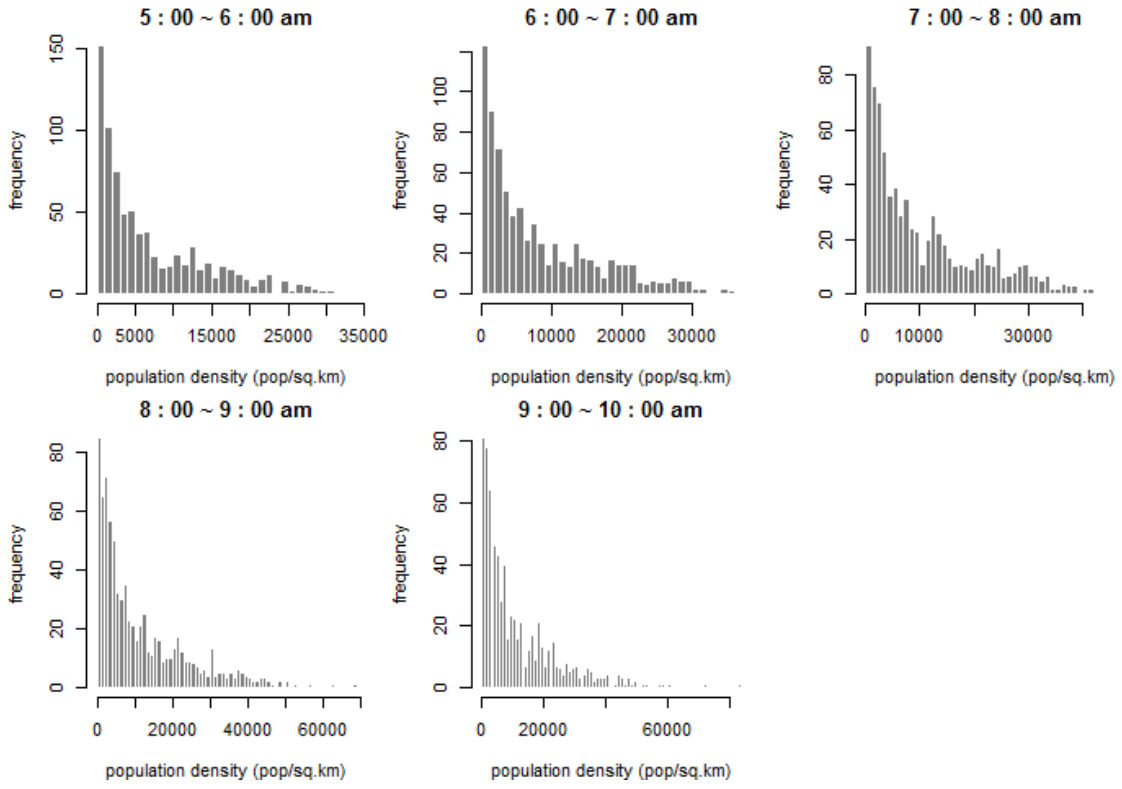

Figure 4: Histograms of Frequency by Hour

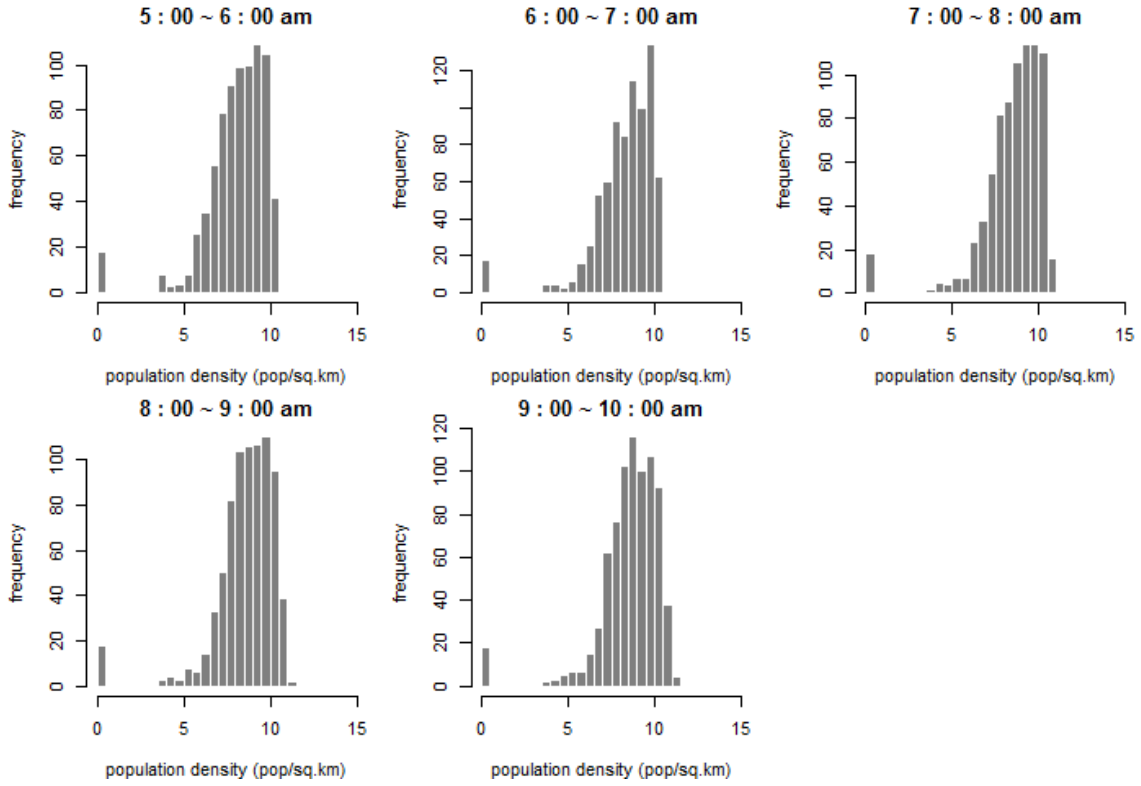

Figure 5: Histograms of log-Frequency by Hour

Image plots are set to be the same scales where dark red color represents low values while bright yellow colors indicate high values. It is clear that bright colors are centered in the middle and the color becomes brighter and brighter by time. That is, more people are gathering to CBD by time.

Histograms show that frequency is highly skewed. However, after log transformation, the histograms are still not desirable (shown in Figure 5). Similarly, boxplot and qqplot also show that variable frequency does not follow normal distribution. After log transformation, it is still not normally distributed. From the histograms, we can see that frequency is in exponential shape with even steeper slope. A Gamma distribution may be more appropriate. A log transformation may not be very helpful.

For the two covariates, they are also exponentially distributed. This indicates that more people are concentrated near metro lines and expressways.

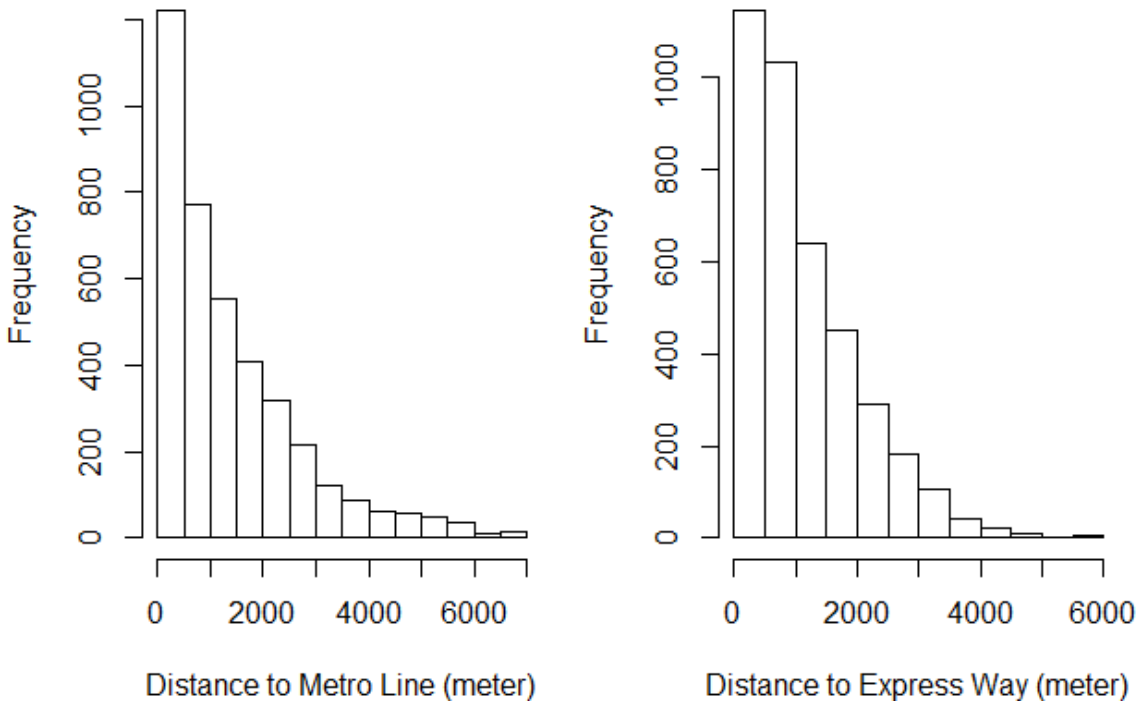

Figure 6: Histograms of Covariates

4.2 Spatial Trend

In order to prepare spatial data for stationarity procss analysis, we need to check the correlation between the response and location. Side by side bar plot is used to demonstrate the relationship between response and location. Population density with respect to latitude and longitude is shown in Figure 7 and Figure 8 respectively. As is shown, there is a curvature relationship between population density and two coordination variables.

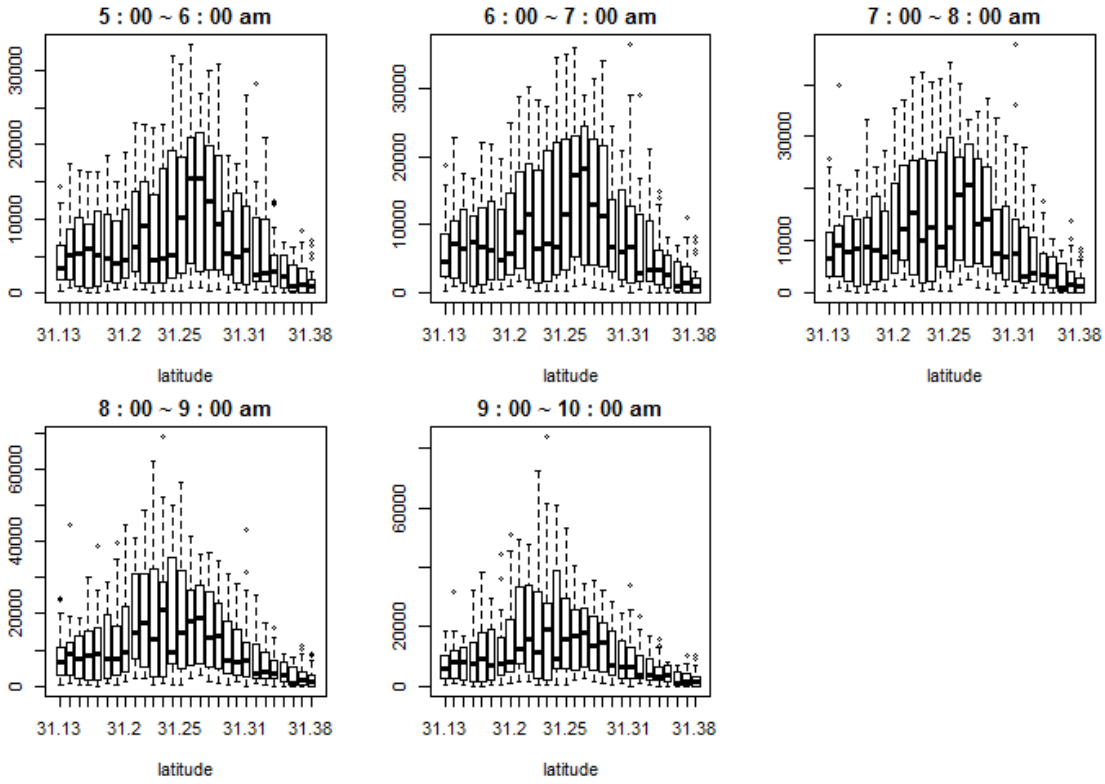

Figure 7: Population Density and Latitude

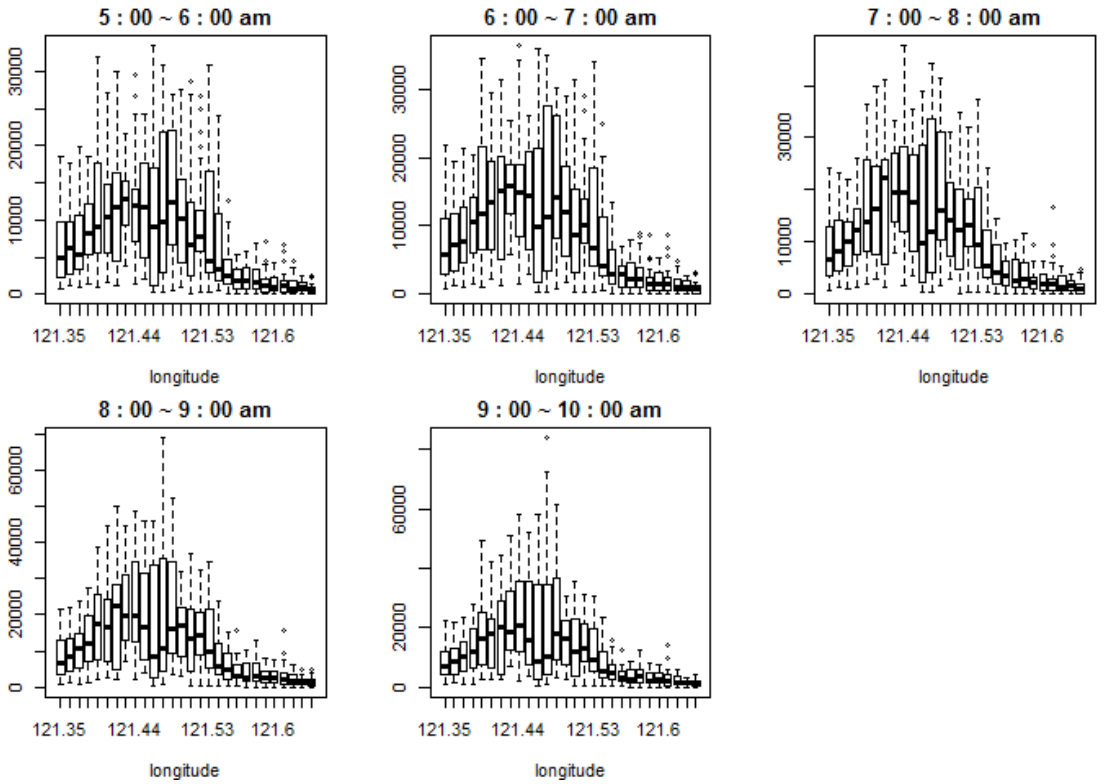

Figure 8: Population Density and Longitude

Therefore, we fit polynomial regression of order two to remove the underlying spatial trend. Naturally, Covariates include latitude, longitude, their quadratic term and interaction. For response, two scenarios are set: in first scenario, we set population density as response; in the other scenario, we set log population density as response. Box-cox transformation is used to help use find best transformation parameter. Figure 9 shows value zero maximize the likelihood across the parameter space, indicating a log transformation may produce better residuals.

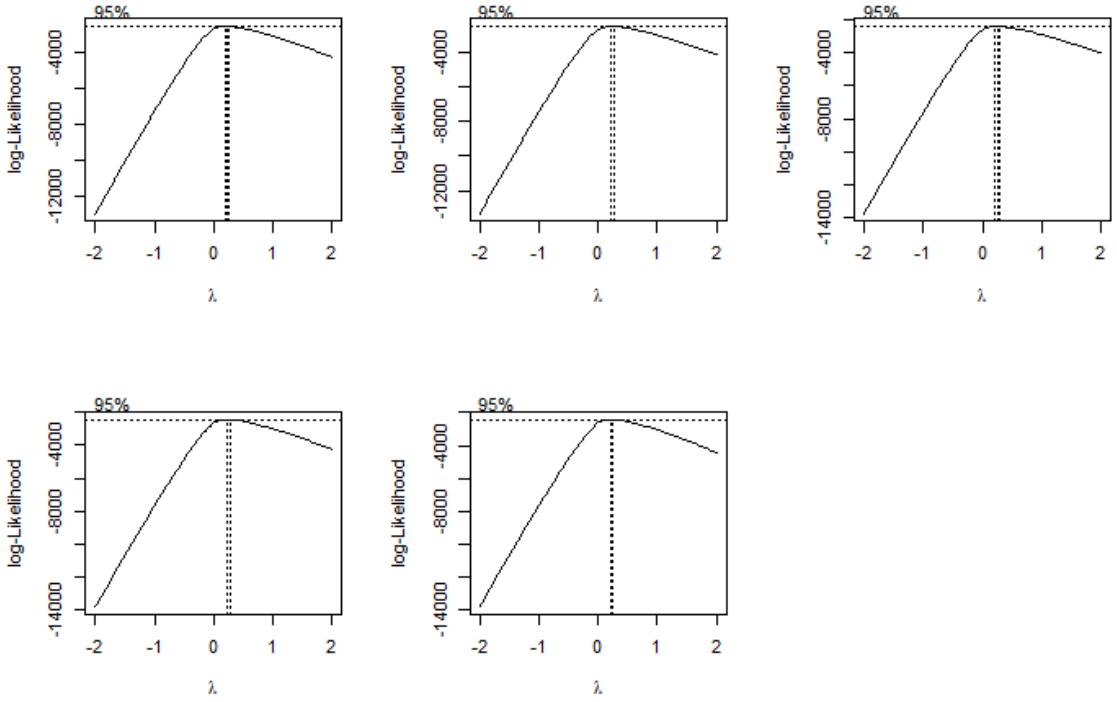

Figure 9: Parameters for Box-Cox Transformation

For both scenarios, stepwise selection is used for variable screening. Under AIC criterion, full model (Model 1) is selected for its best performance.

$$Y(s_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + \epsilon_i \tag{1}$$

where,
$Y(s_i)$ is population density at location s_i;
x_{1i} is latitude of location s_i;
x_{2i} is longitude of location s_i;

For each hour, we remove the underlying spatial trend using polynomial regression model 1 fitted above. By comparing two scenarios, we find that there is no observable improvement in detrending for log scale scenario while the improvement in residual for original scale scenario is obvious. As a result, we adopt the polynomial regression model in scenario one: population density as response. Part of our fitting results is show below as parameter estimates for $5:00 - 6:00$ am:

$\hat{\beta}_0 = -4.524 * 10^9$;
$\hat{\beta}_1 = 2.018 * 10^7$;
$\hat{\beta}_2 = 6.932 * 10^7$;
$\hat{\beta}_3 = -5.698 * 10^5$;
$\hat{\beta}_4 = -3.018 * 10^5$;
$\hat{\beta}_5 = 1.271 * 10^5$.
Estimates for other time periods are shown in Appendices.

As is displayed in Figure 10 and Figure 11 , spatial pattern removed after detrending. Residuals for 5:00 to 6:00 am fall in a constant band within -10000 and 10000. No observable spatial trend is discovered after removing the underlying process. Similar results for other hours are shown in plots in Appendices.

Thus,instead of using population density, residuals for each location produced in Model 1 are used to conduct spatial analysis in the following sections.

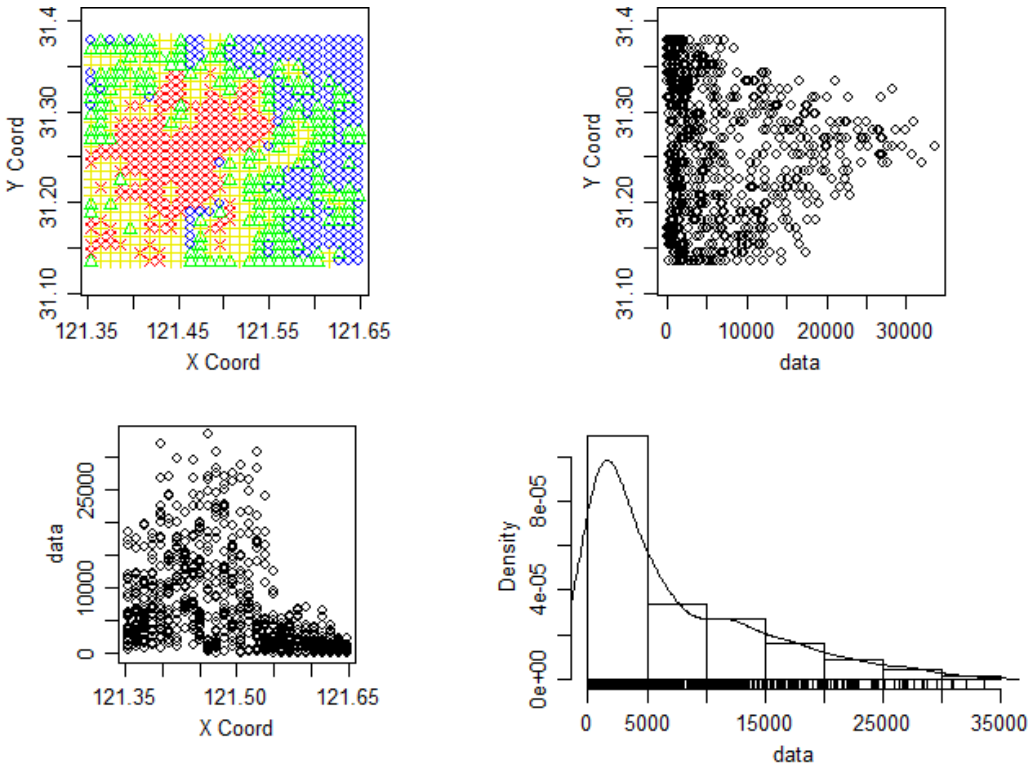

Figure 10: Exploratory Plot before Detrending(5:00-6:00 am)

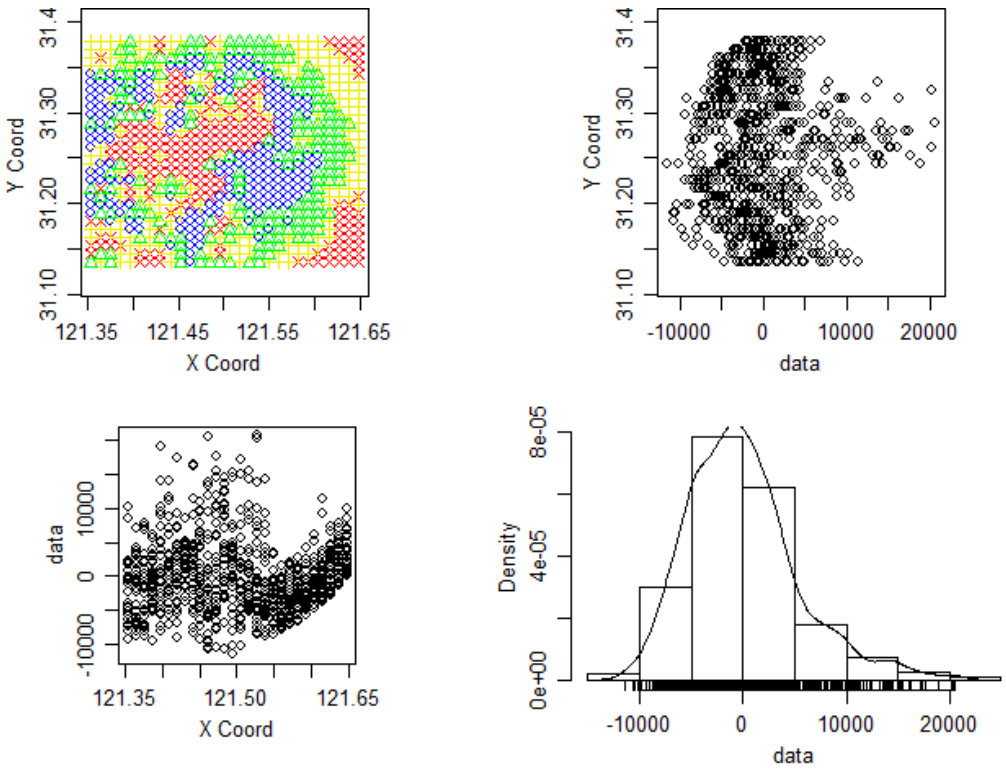

Figure 11: Exploratory Plot after Detrending(5:00-6:00 am)

4.3 Isotropy

Isotropy means the covariance function of the spatial random variable depends on the displacement vector only through its length. This is always a desirable property as it simplifies the covariance structure and we can express the covariance structure as a function of distance.

To check for isotropy assumption, two widely used methods are to look at the empirical semivariogram contour plots and directional semivariance plots. First we start with directional semivariance plot shown in Figure 12. The plots show that semivariances for different direction are close to each other except the tail part where sample size is small, indicating there is no significant difference in semivariance among all the direction plotted. For complementary, we further check the empirical semivariogram contour plots shown in Figure 13. The contour plot shows unique peak in the center of study area, outlined by circular contour lines which indicating an acceptable isotropy situation.

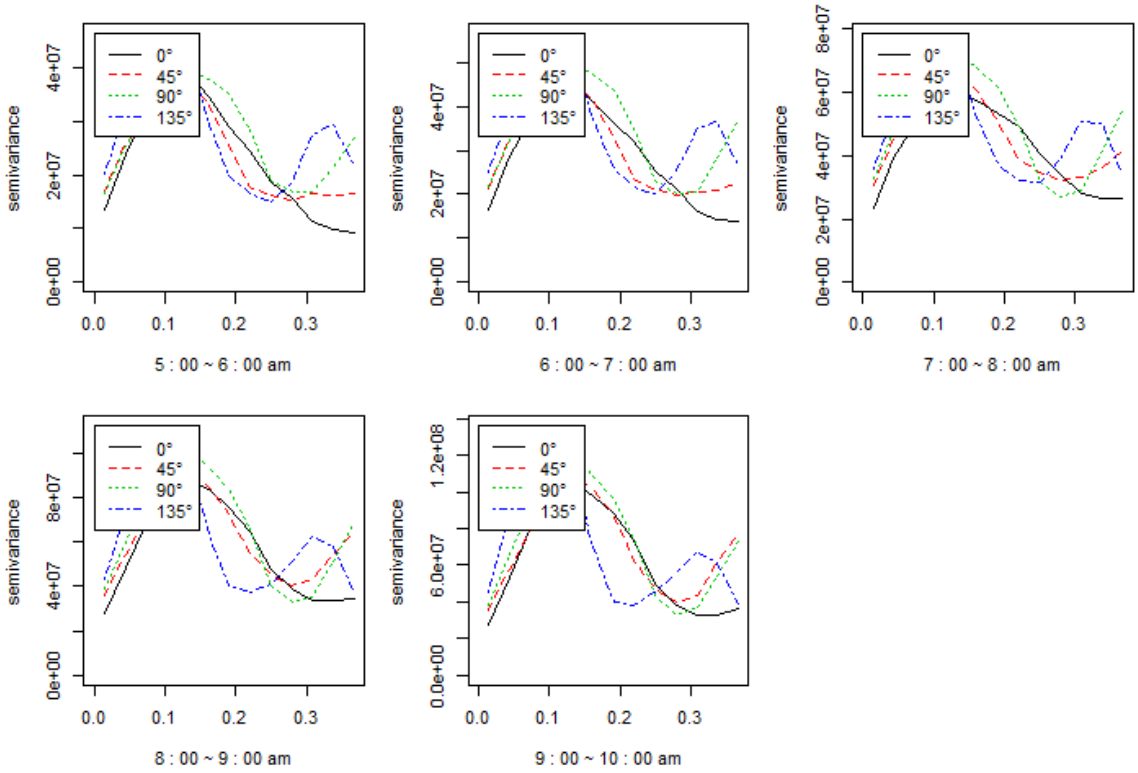

Figure 12: Directional Semivariance Plots

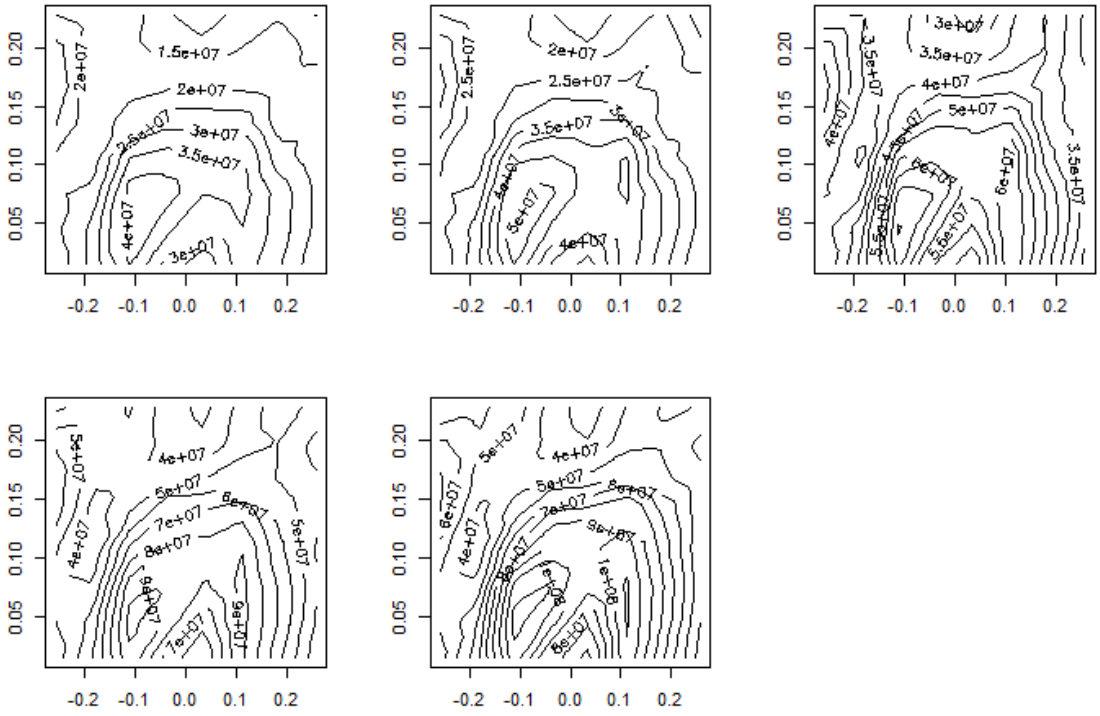

Figure 13: Empirical Semivariogram Contour plot

4.4 Model Fitting

All preliminary work has been done. The next step is analyzing the processed data. Again, we start from visualization to study the intrinsic properties of the data.

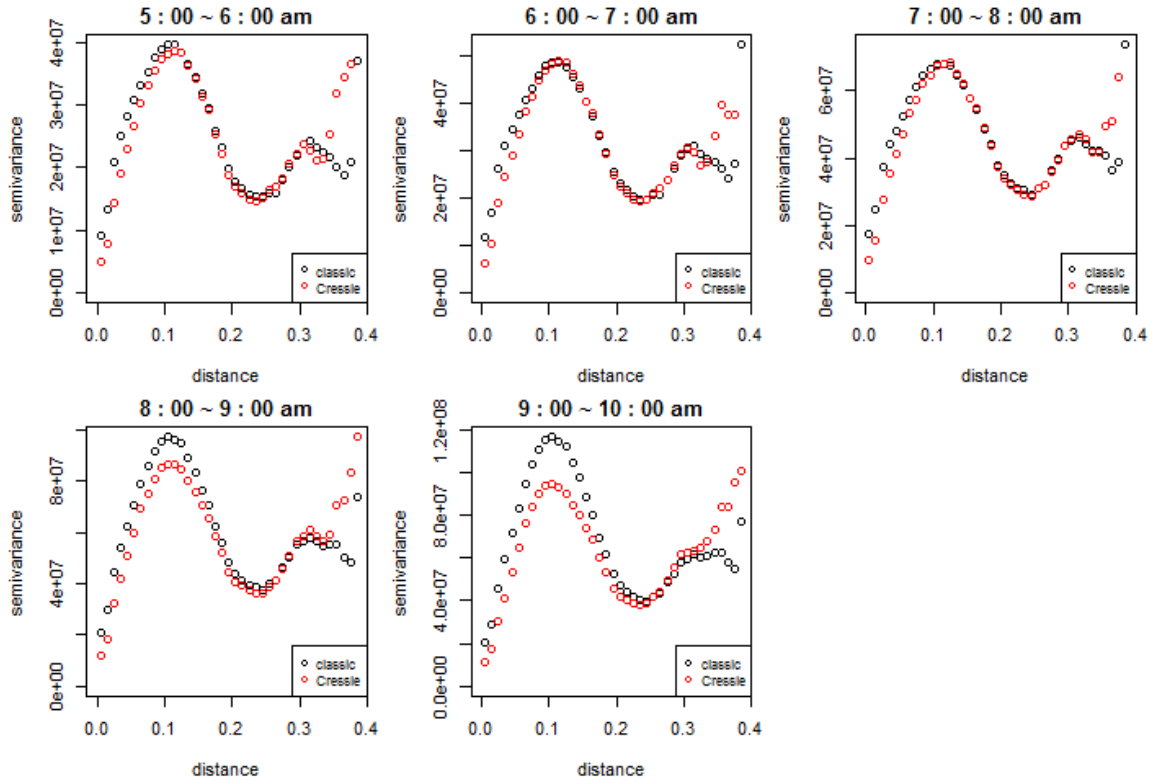

Figure 14: Semivariogram with Classical and Cressie Weight

Figure 14 shows the semivariogram by time. There is obvious difference between the classical and Cressie weights, especially in 8-9 and 9-10 time period. As the scale is quite large, the difference is severe. Since Cressie weight is more robust, which can also be seen from plots, Cressie weight is used in the analysis.

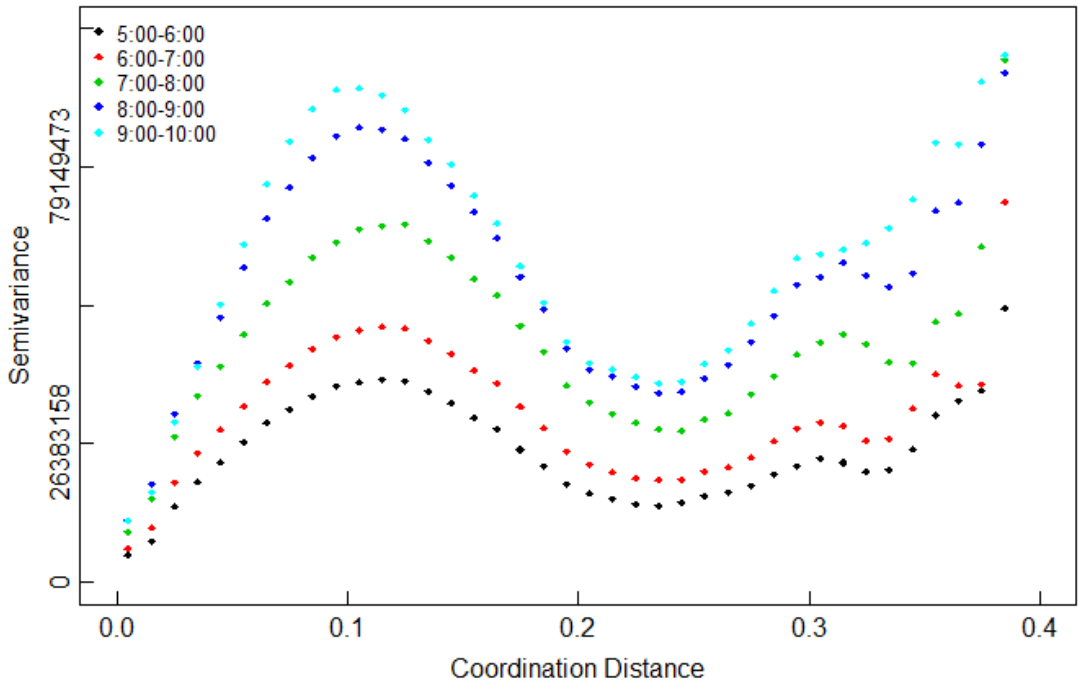

Figure 15: Semivariogram with All Time in One

Figure 15 further convinces our induction in Section 3 in that the semivariogram becomes decreasing after 0.1. We select 0.15 as our largest distance in further analysis as 0.15 corresponding to a roughly 14km. In addition, Figure 15 also shows that the variance is increasing by time. The curves are more steady in earlier times while change severely at later times. This implies a trend from homogeneity to heterogeneity.

4.4.1 Semivario Fitting

Next, we try to capture the behavior of semivariogram within 0.15 distance by fitting semivariogram models. Cressie weight is applied while 8 covariance matrices are considered: spherical, exponential, Gaussian, cubic, Matern, circular, power and power.exponential. The corresponding estimates of parameters with residual sum of squares are listed in Tables in Appendix B. Table 2 lists the residual sum of squares of the eight models in different time slots. The last column calculates the mean residual of different time for each model. Based on the result, we can see the best model with smallest residual sum of squares is the Spherical model. The results are shown in Table 10. The practical range is around 0.1, which is smaller than 0.15, our choice of maximum distance.

4.4.2 Likelihood Fitting

Likelihood fitting method is conducted in a similar way. Five covariance models: spherical, exponential, Gaussian, cubic and Matern are considered. Estimations with

Table 2: Comparison Between Different Semivariogram Models

	5	6	7	8	9	mean
Spherical	1345.8285	1492.5117	1584.2604	1087.8530	1313.3275	1364.7562
Exponential	1784.8855	1777.4990	1737.4838	1778.2459	2688.8343	1953.3897
Gaussian	1585.4945	1768.0713	1850.2990	1292.0789	1369.6244	1573.1136
Cubic	1564.8268	1744.1053	1826.2688	1236.8298	1278.8814	1530.1824
Matern	1784.8855	1777.4990	1737.4838	1778.2459	2688.8343	1953.3897
Circular	1388.7989	1544.0794	1635.9566	1082.5673	1200.9239	1370.4652
Power	4369.2902	4148.1227	3914.7066	4643.6822	6112.1596	4637.5923
Powered.exponential	3177.1008	3017.0725	2784.1867	3302.9481	4695.4287	3395.3474

AIC and BIC for each model are listed in tables in Appendix B. Table 3 summarizes the AIC and BIC for each model at different time with a mean at the last column. The smallest AIC and BIC models are exponential and Matern models, which have exactly the same results. The practical range is also around 0.1, less than 0.15.

Table 3: Comparison Between Different Likelihood Models

	5	6	7	8	9	mean
Spherical AIC	4132.6943	4342.5478	4689.4731	4830.2513	4771.8406	4553.3614
Exponential AIC	4108.0388	4315.9060	4659.6106	4805.0298	4756.6708	4529.0512
Gaussian AIC	4859.8724	5033.1905	5318.4163	5550.9594	5661.5423	5284.7962
Cubic AIC	4123.6046	4336.6447	4684.5895	4854.4454	4831.2214	4566.1011
Matern AIC	4108.0388	4315.9060	4659.6106	4805.0298	4756.6708	4529.0512
Spherical BIC	4151.3520	4361.2055	4708.1308	4848.9089	4790.4983	4572.0191
Exponential BIC	4126.6964	4334.5636	4678.2683	4823.6874	4775.3285	4547.7088
Gaussian BIC	4878.5301	5051.8482	5337.0739	5569.6170	5680.1999	5303.4538
Cubic BIC	4142.2623	4355.3024	4703.2471	4873.1031	4849.8790	4584.7588
Matern BIC	4126.6964	4334.5636	4678.2683	4823.6874	4775.3285	4547.7088

4.4.3 Effects of Covariates

Afterwards, we study the effects of covariates on frequency. Figure 16 and Figure 17 show the relationship of frequency on the two covariates over time.

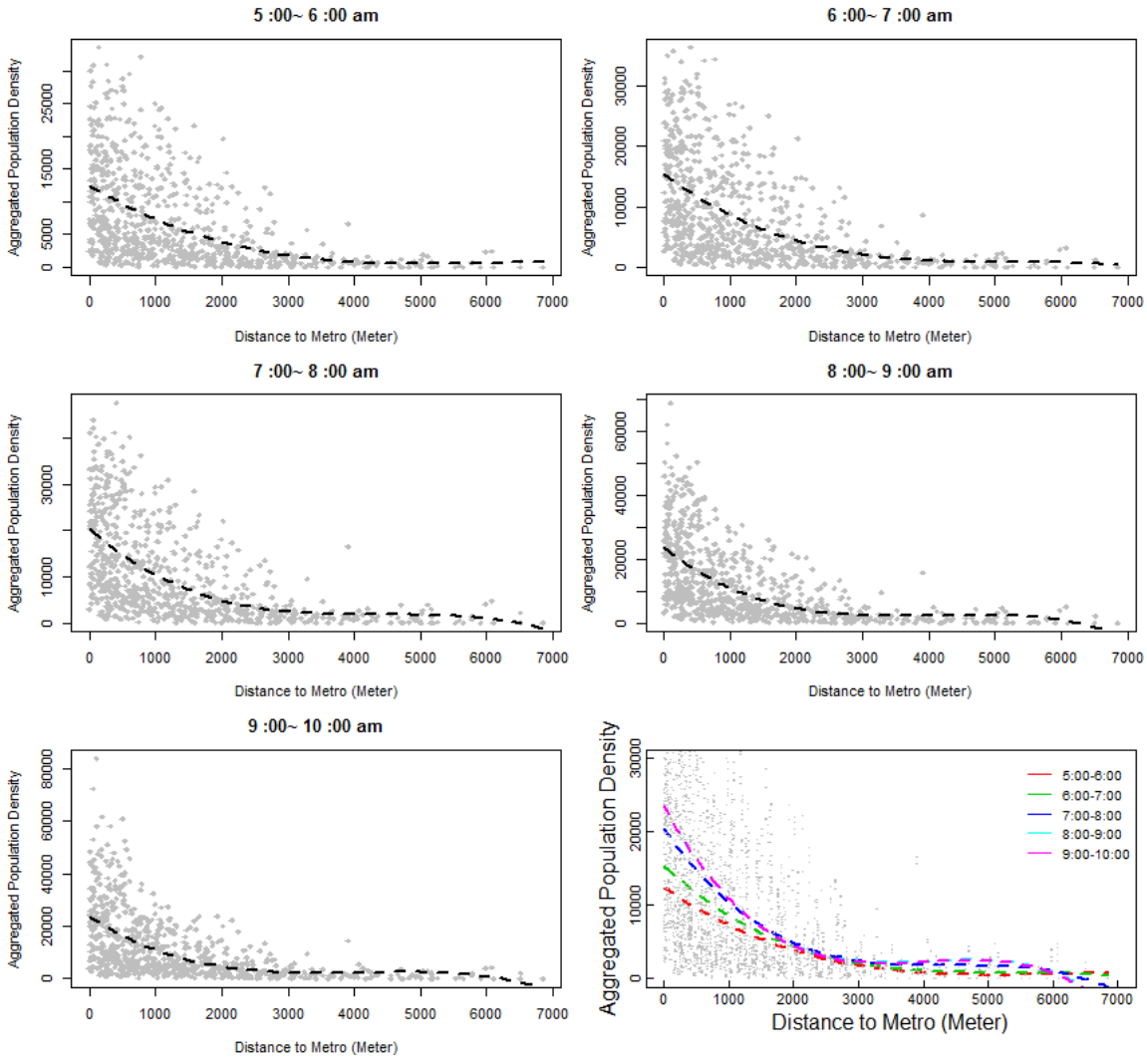

Figure 16: Relationship of Frequency on Distance2Metro

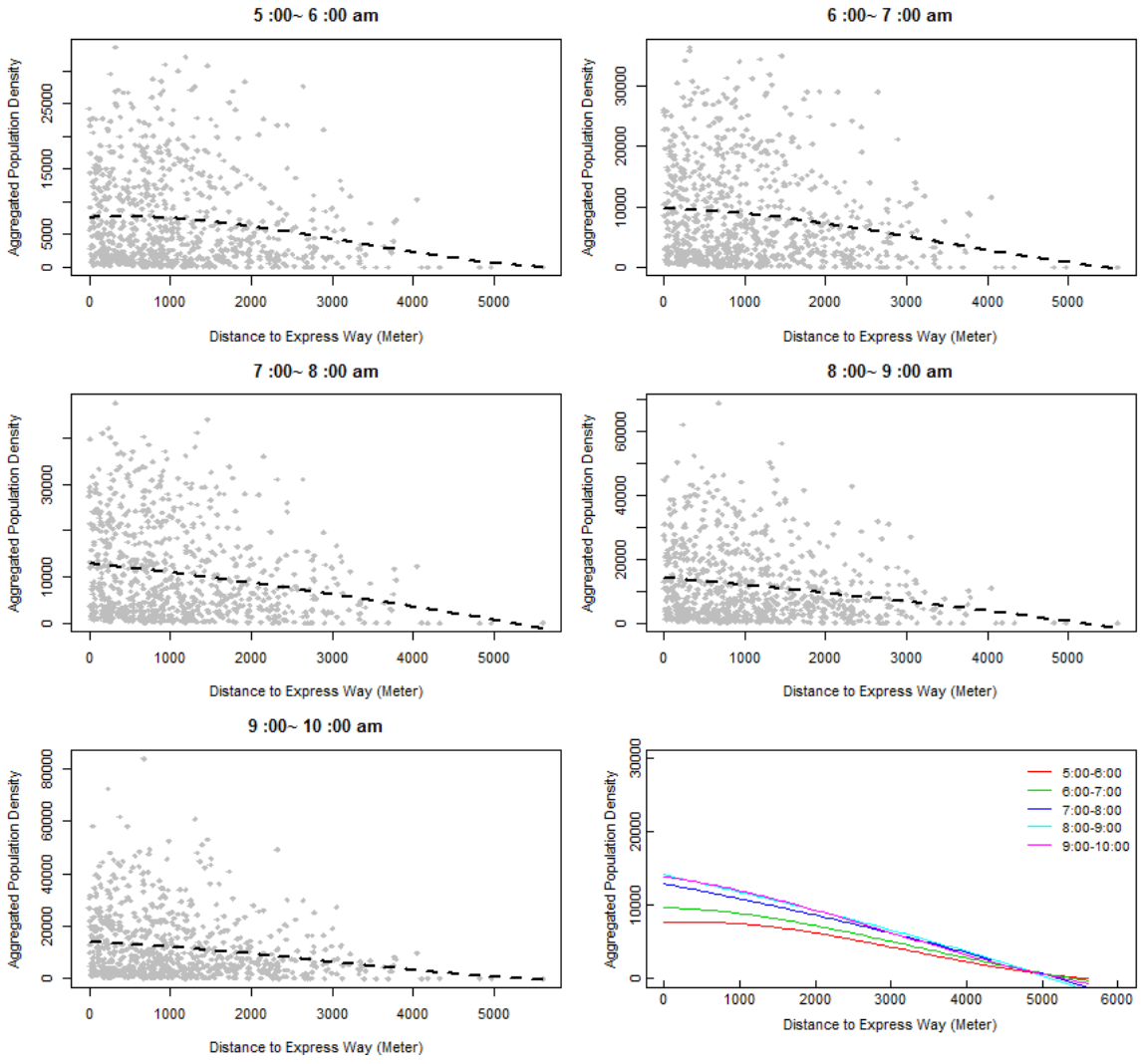

Figure 17: Relationship of Frequency on Distance2Expressway

Distances to transportation corridors have impact on population density with the smaller the distance, the larger the density. Also, as time increases, the population density also increases and the increase is more concentrated with shorter distances. Thus we conclude that the closer to transportation corridors, the more people. In addition, as time goes by, more people are gathering towards the metros or expressways.

As shown in Section 4.1, the frequency variable does not follow normal distribution, so a generalized model is considered. The histogram in Figure 4 shows that a Gamma distribution seems appropriate for frequency. Also, the frequency, or density variable can be treated as count data due to our data aggregation method with disperse variance. Thus negative binomial family is also considered. For both with and without spatial effects model, we start with full quadratic terms. AIC is used in model selection. For GLM model with negative binomial family, AIC is not available. Thus residual sum of squares is compared to GLM model with Gamma family and the Gamma family has a much smaller residual sum of squares. After all comparisons, the best model is GAM model with spatial effect using Gamma family. Results are shown in Table 4. Results for other models are shown in Appendix B.

Table 4: Summary for Gamma GAM Spatial Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.00	0.00	44.06	0.00
D2Metro	0.00	0.00	1.51	0.13
D2Road	-0.00	0.00	-2.57	0.01
I(D2Metro^2)	0.00	0.00	3.88	0.00
D2Metro:D2Road	0.00	0.00	2.30	0.02

Histograms show that frequency is highly skewed. However, after log transformation, the histograms are still not desirable (shown in Figure 5). Similarly, boxplot and qqplot also show that variable frequency does not follow normal distribution. After log transformation, it is still not normally distributed.

Most of the coefficients are significant, further indicating that the covariates have impact on population density. The impact is huge since the parameter estimations are very close to zero because Gamma family uses inverse link. Results from model further prove previous conclusions from visualization.

4.5 Prediction

For visualization, we only present a comparison of true image with predicted images at 5:00 - 6:00 am, shown in Figure 18. Residual sum of squares are compared at all time periods (Figure 19).

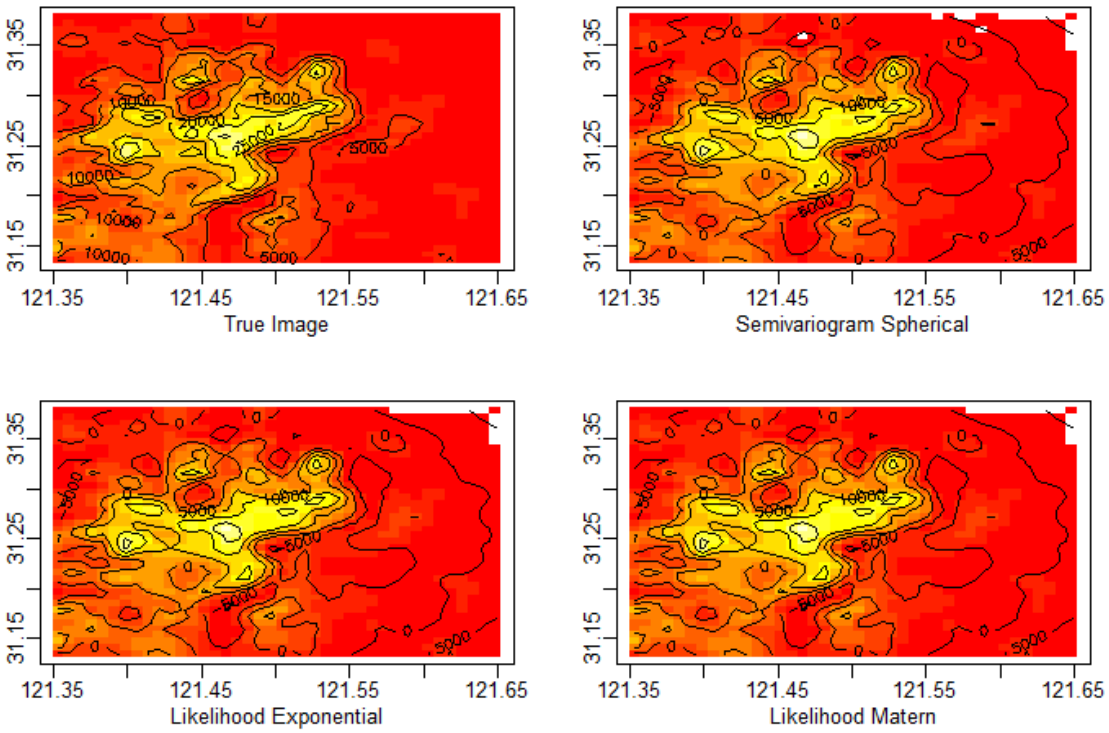

Figure 18: Comparison between True Model and Predicted Models

	Spherical	Exponential	Matern
5	483239486	441347900	441347900
6	640487307	585196252	585196252
7	1018205620	917097306	917097306
8	1172933529	1074494482	1074494482
9	1071913221	980410306	980410306

Figure 19: Summary of Residual Sum of Squares between True Model and Predicted Models

Although in each individual fitting, empirical methods fit better than likelihood fitting (see Figure 34 and Figure 43 in Appendix), for predicted models, the likelihood models are better than empirical fittings.

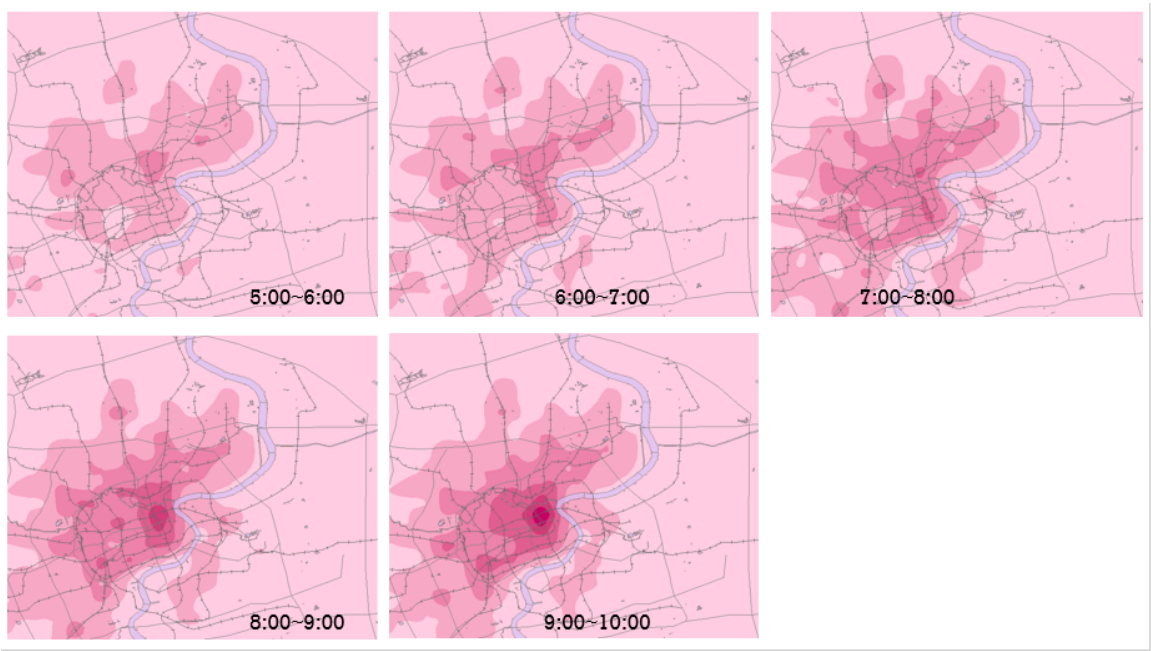

Figure 20: Predictions with Exponential Likelihood Models by Hour

Figure 20 displays predictions over time from likelihood fitted models with exponential covariance. It is clear that population movement is along transportation corridors.

5 Conclusion

Nowadays, urban planners are facing with more challenges to adapt to rapid growth of modern metropolitan and urban dynamics. Effectively responding to urban problems such as congestion and stampede is more emphasized than ever before. Cellular data provides a promising solution for such problems for its effectiveness and high resolution. In this paper, we propose a method to demonstrate the spatial and temporal characteristics of population movement both quantitatively and qualitatively, based on cellular data.

In this study, we aggregated our data on an hourly basis, and this avoids us to obtain any further conclusion of higher accuracy in terms of time and population density. In the future, we plan to aggregate the data with higher resolution to make more detailed conclusion.

This paper analyzes population movement in Shanghai during morning peak hours on a typical weekday using point-referenced data models in spatial analysis. Based on our study, especially the prediction plots, we have several findings:

- People are gathering towards CBD area by time;
- As time goes by, population density becomes more various;
- The closer to transportation corridors, the more people gathering;
- Increase in population is larger near metros or expressways;

- Population movement in along transportation corridors.

Due to time limitation, this report still has room for improvement.

- First, we also did areal analyses but did not finish it. We could conducted areal analysis and then compare the results with geostatistical analysis.
- Second, we can repeat the whole process for residuals from model with covariates. That is, the detrend model can include covariates.
- Third, we should also consider Bayesian models and then compare between the results.
- Fourth, a spatio-temporal model may be considered instead of looking at the trend by time from plots.
- Fifth, When making predictions, a leave-one-out cross validation may be applied to get a more accurate prediction.

6 Reference

World Bank. Urban population, 2015.
`http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS`

Mindali O, Raveh A, Salomon I. Urban density and energy consumption: a new look at old statistics [J]. Transportation Research Part A: Policy and Practice, 2004, 38(2): 143-162.

Foley D L. The daily movement of population into central business districts [J]. American Sociological Review, 1952, 17(5): 538-543.

Willett K. Stuck in traffic and stuck for solutions: Brisbanes congestion crisis[C]. Australian Institute of Planning and Management (Qld) Seminar, Brisbane, QLD, 2006.

Olayiwola K O, Olaseni A M, Fashina O. Traffic Congestion Problems in Central Business District (CBD) Ikeja, Lagos Metropolis, Nigeria[J]. Journal of Research on Humanities and Social Sciences, 2014, 4(1): 23-32.

Thriault, Marius, et al. "Modelling commuter trip length and duration within GIS: Application to an OD survey." Journal for Geographic Information and Decision Analysis 3.1 (1999): 41-55.

Dong, Honghui, et al. "Urban traffic commuting analysis based on mobile phone data." 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, 2014.

Center Transportation Studies. "Final Evaluation Report for the CAPITAL-ITS Operational Test and Demonstration Program." University of Maryland College Park

(1997).

Fontaine, Michael D., Anjani P. Yakkala, and Brian Lee Smith. Probe sampling strategies for traffic monitoring systems based on wireless location technology. No. FHWA/VTRC 07-CR12. 2007.

Liu, H. X., Danczyk, A., Brewer, R., Starr, R. (2008). Evaluation of cell phone traffic data in Minnesota. TRB 87th annual meeting compendium of papers CD-ROM.

Steenbruggen, John, et al. "Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities." GeoJournal 78.2 (2013): 223-243.

A Appendices – Plots

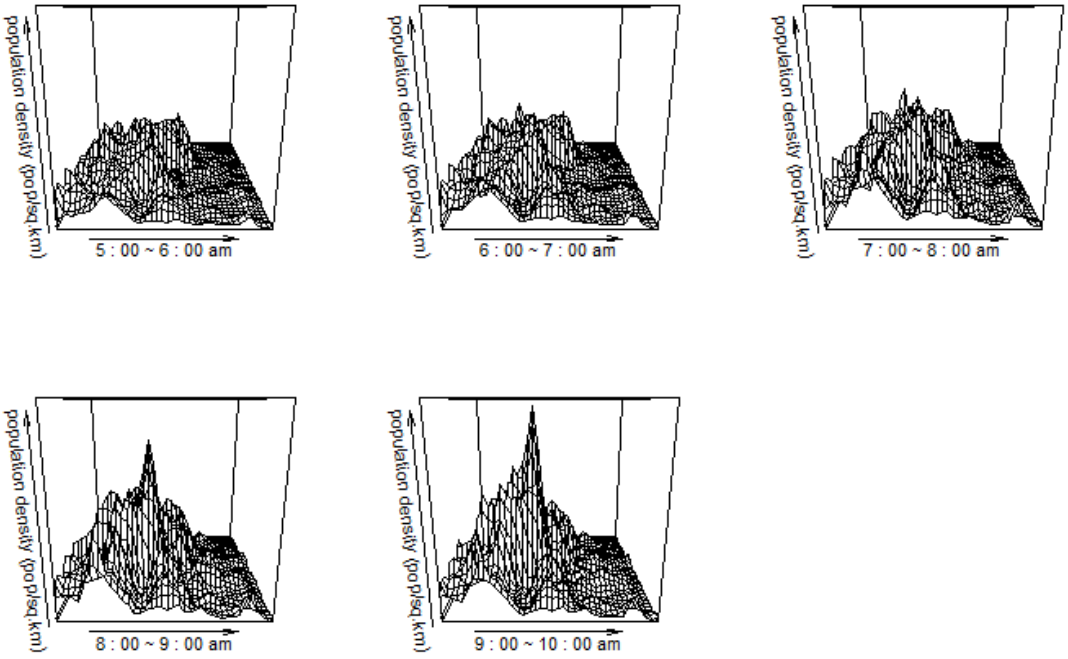

Figure 21: Perspective Plot of Frequency by Hour

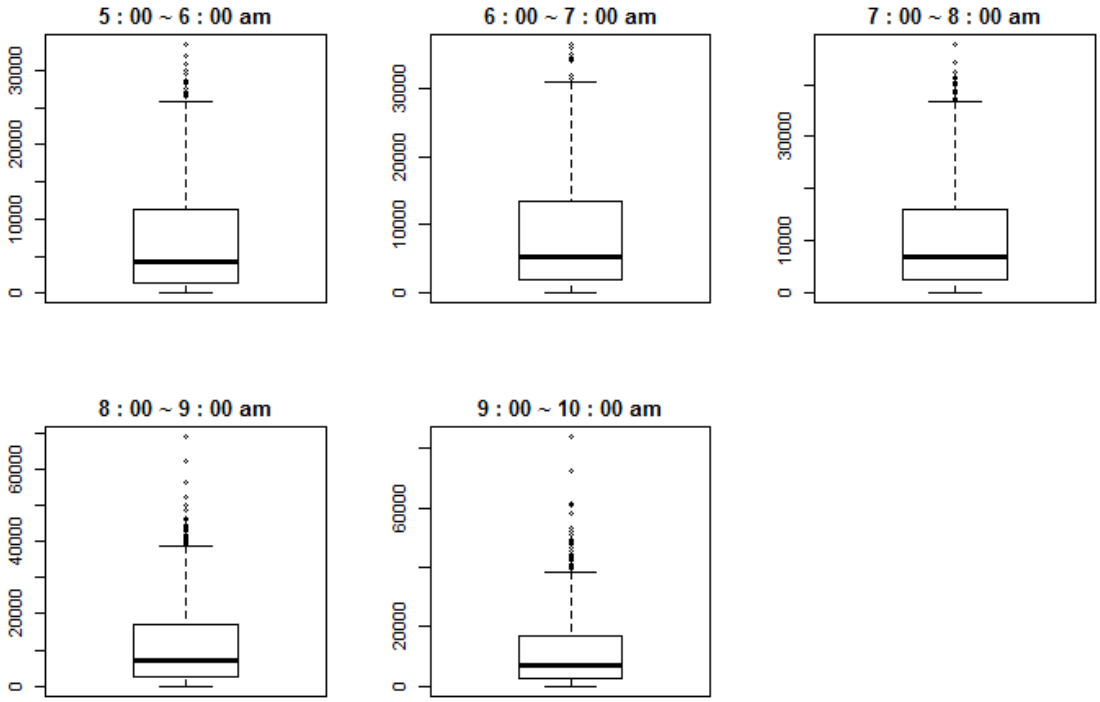

Figure 22: Boxplots of Frequency by Hour

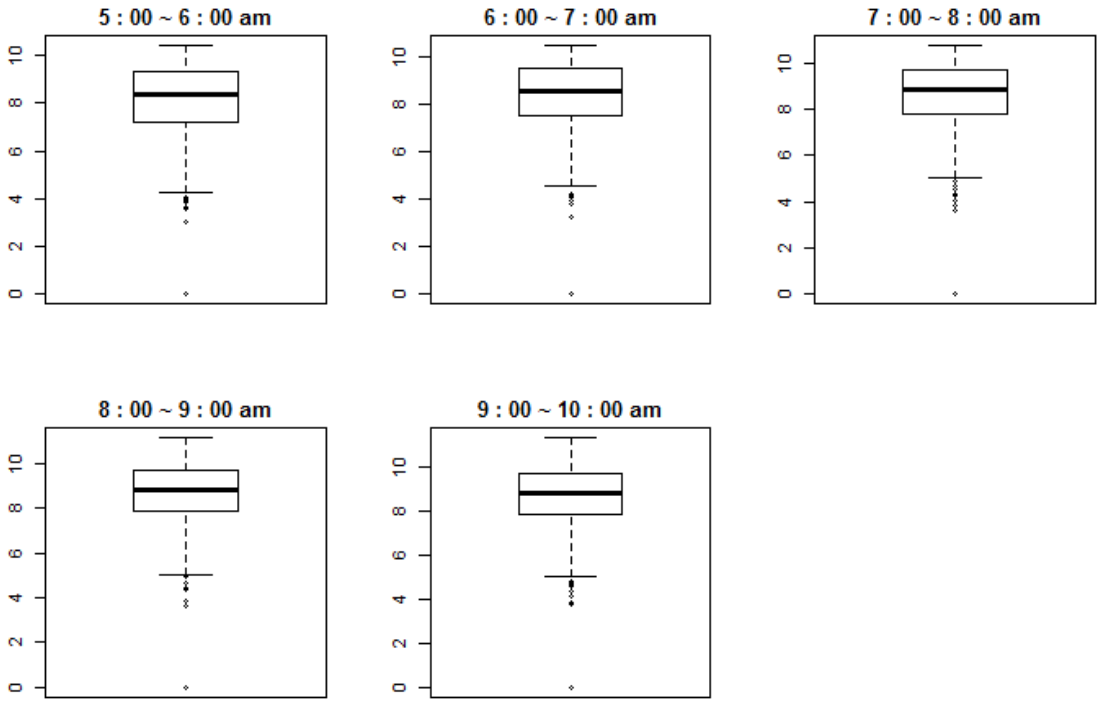

Figure 23: Boxplots of log-Frequency by Hour

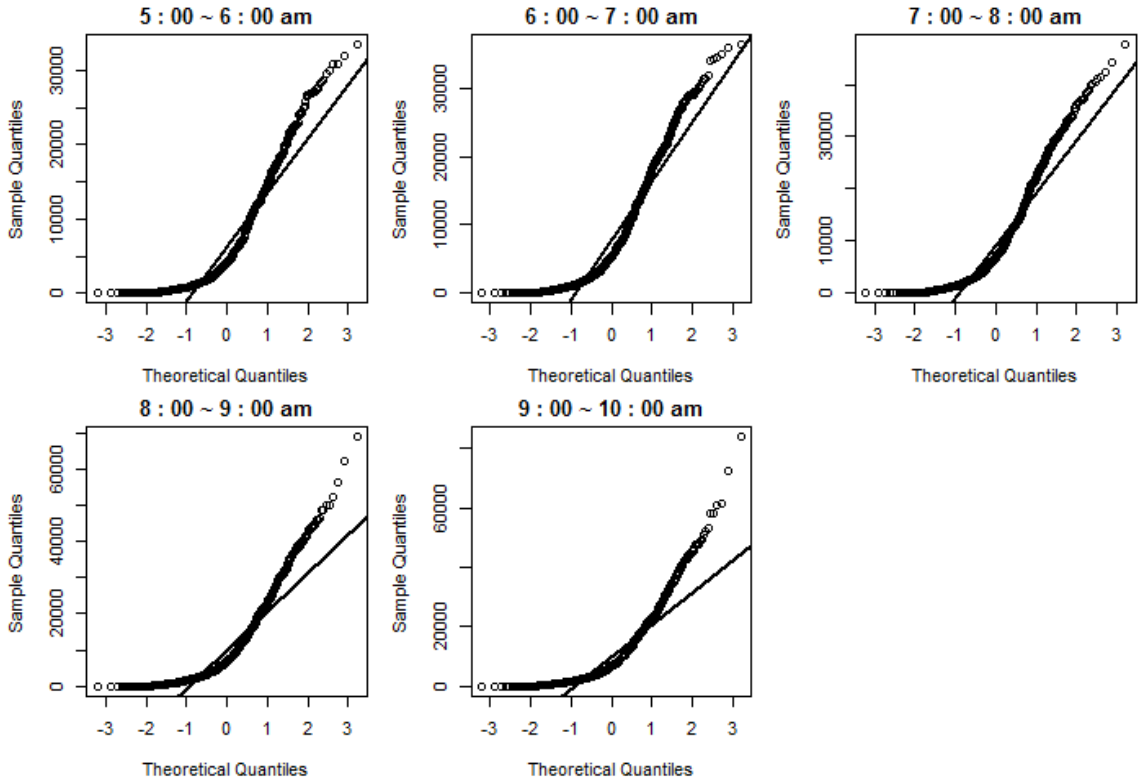

Figure 24: QQ-Plots of Frequency by Hour

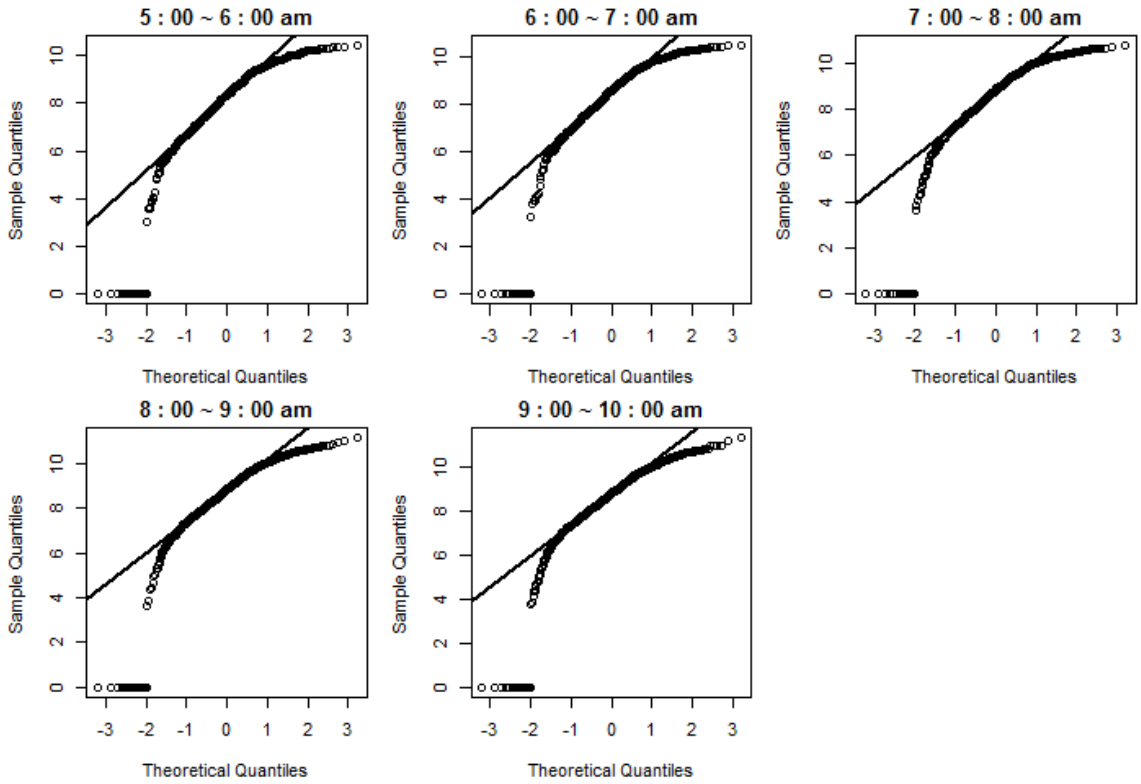

Figure 25: QQ-Plots of log-Frequency by Hour

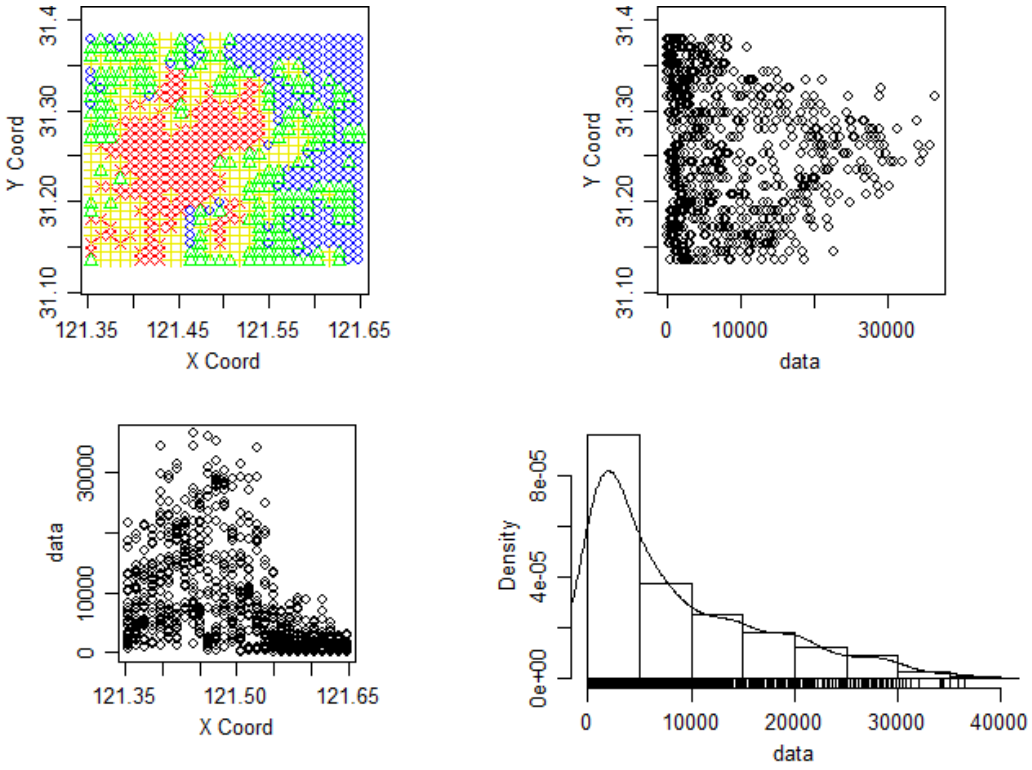

Figure 26: Exploratory Plot before Detrending(6:00-7:00 am)

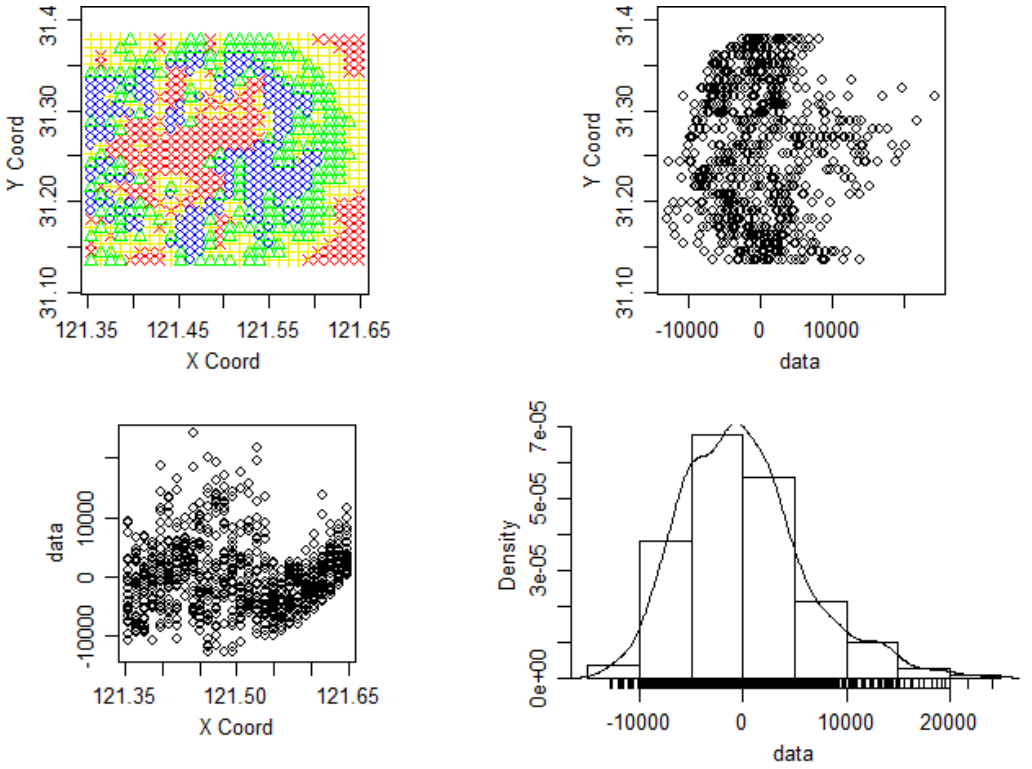

Figure 27: Exploratory Plot after Detrending(6:00-7:00 am)

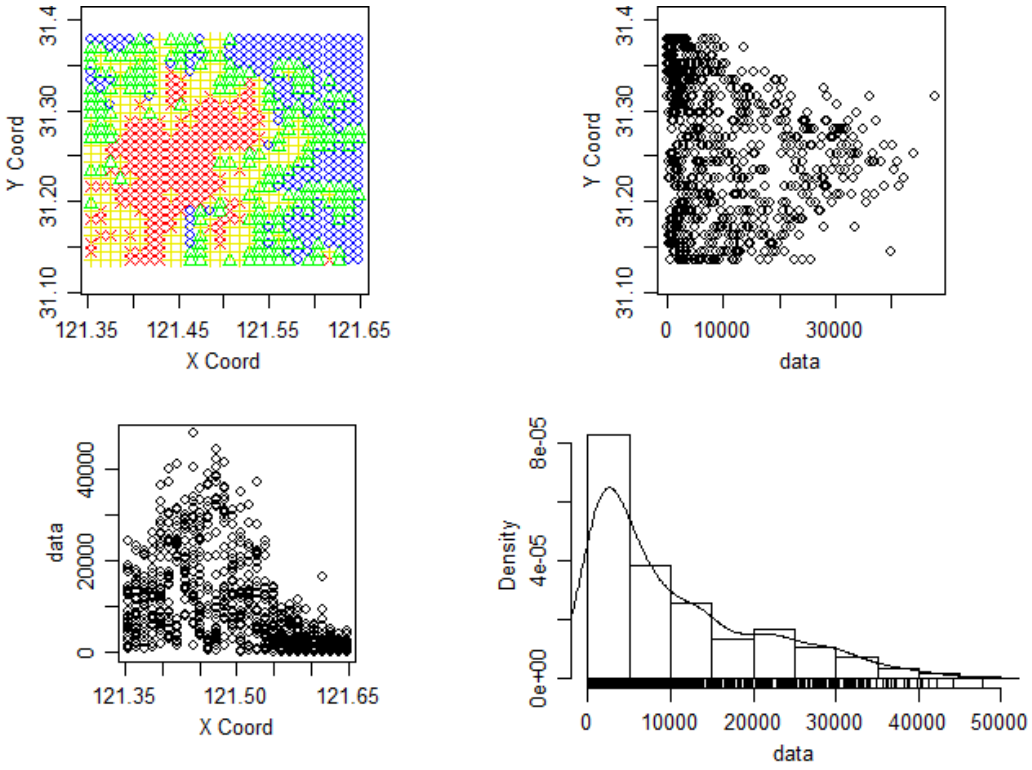

Figure 28: Exploratory Plot before Detrending(7:00-8:00 am)

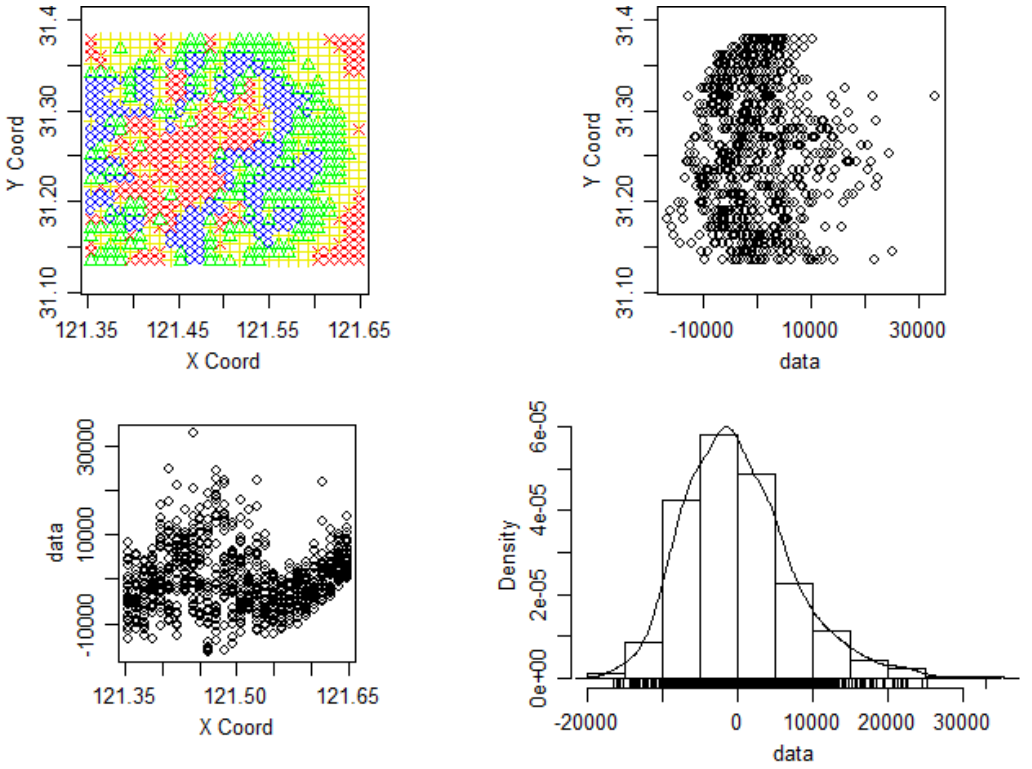

Figure 29: Exploratory Plot after Detrending(7:00-8:00 am)

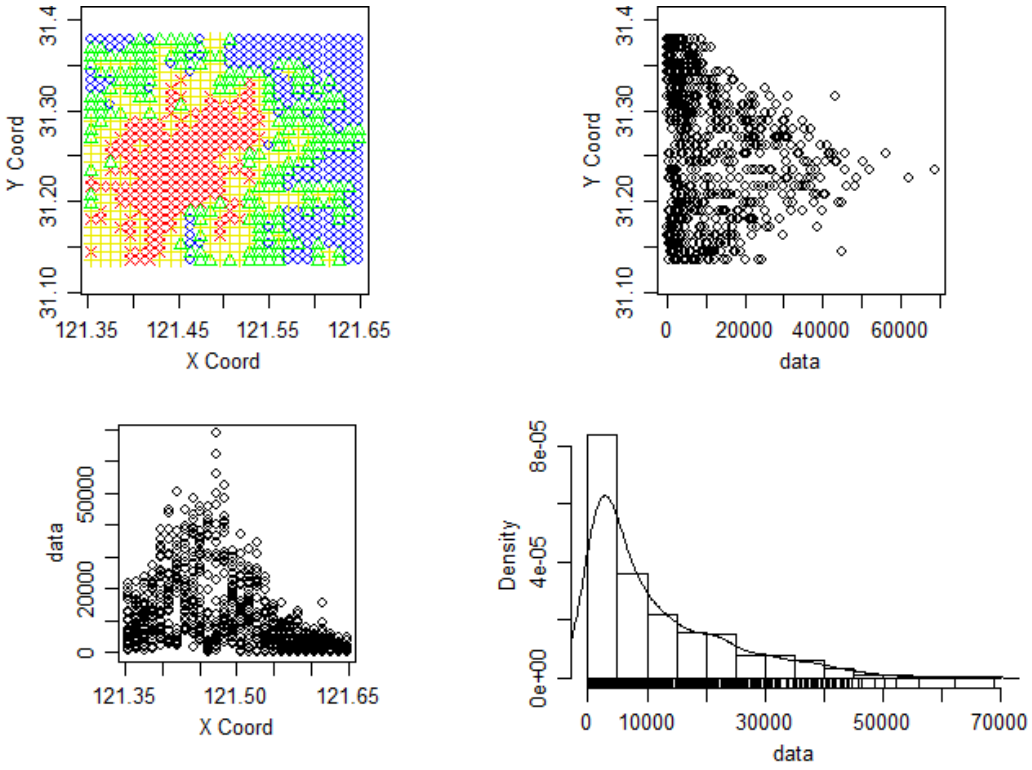

Figure 30: Exploratory Plot before Detrending(8:00-9:00 am)

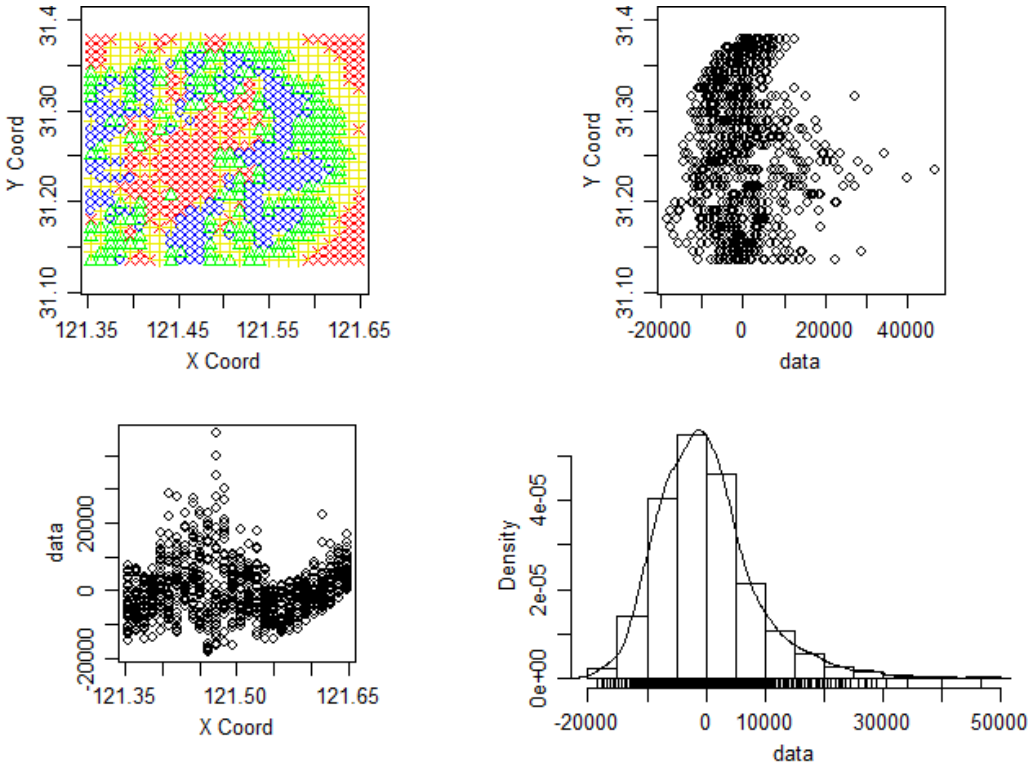

Figure 31: Exploratory Plot after Detrending(8:00-9:00 am)

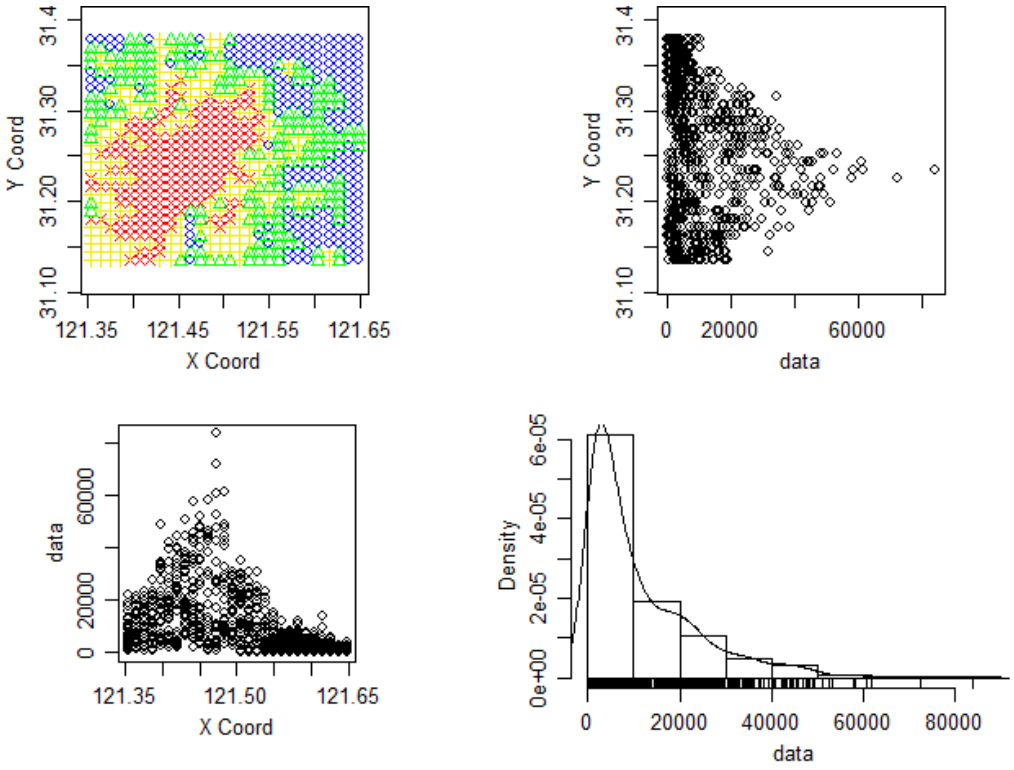

Figure 32: Exploratory Plot before Detrending(9:00-10:00 am)

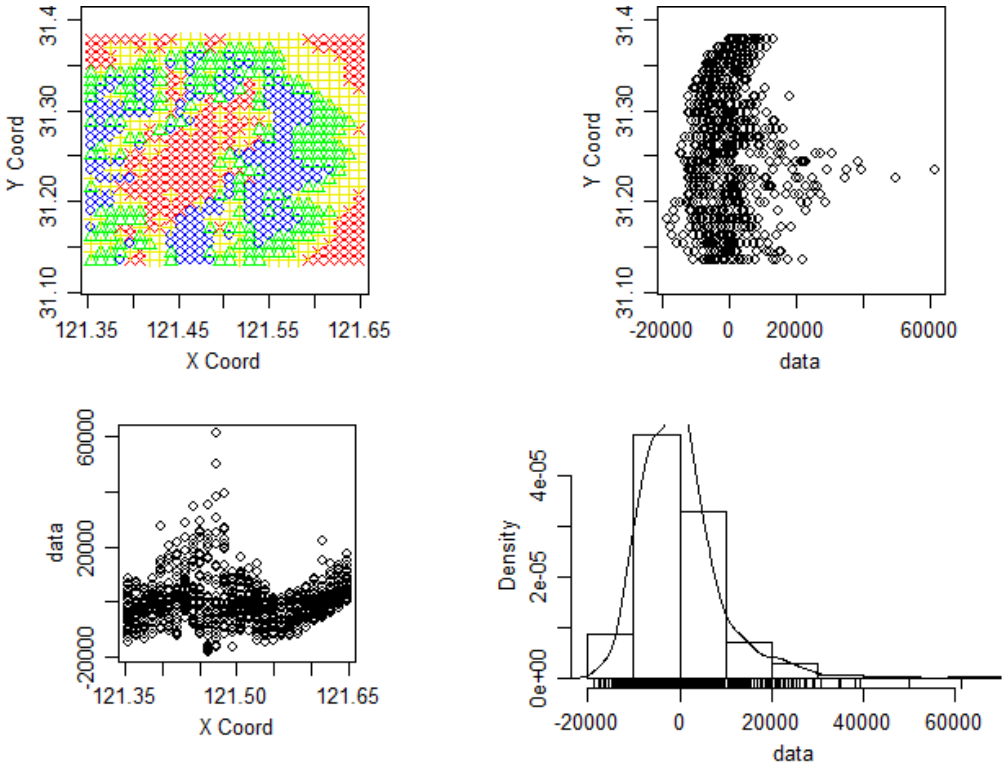

Figure 33: Exploratory Plot after Detrending(9:00-10:00 am)

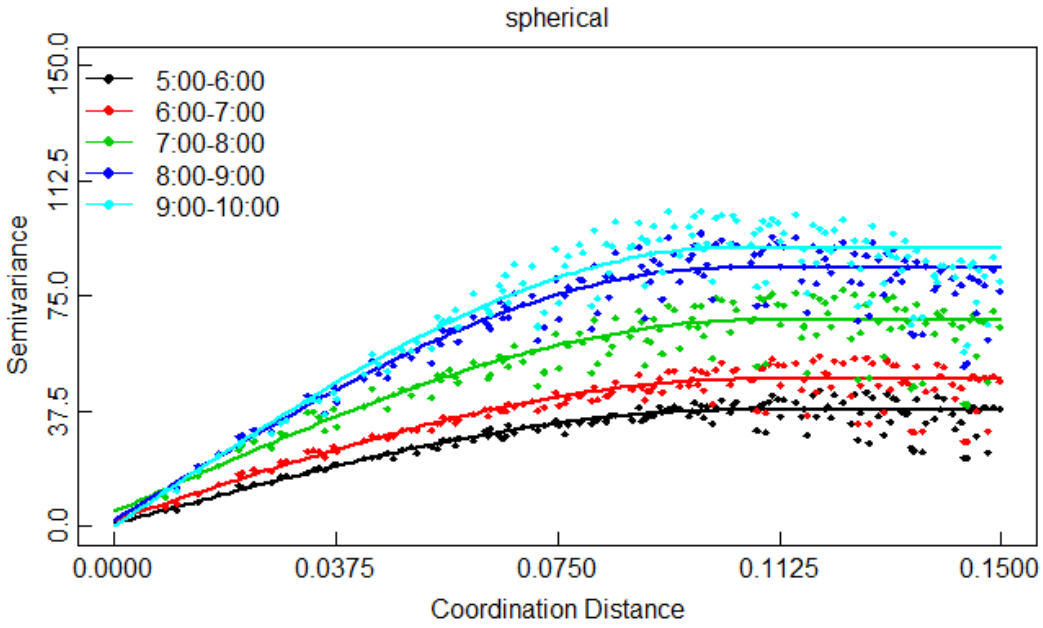

Figure 34: Empirical Fitted Model with Spherical Covariance

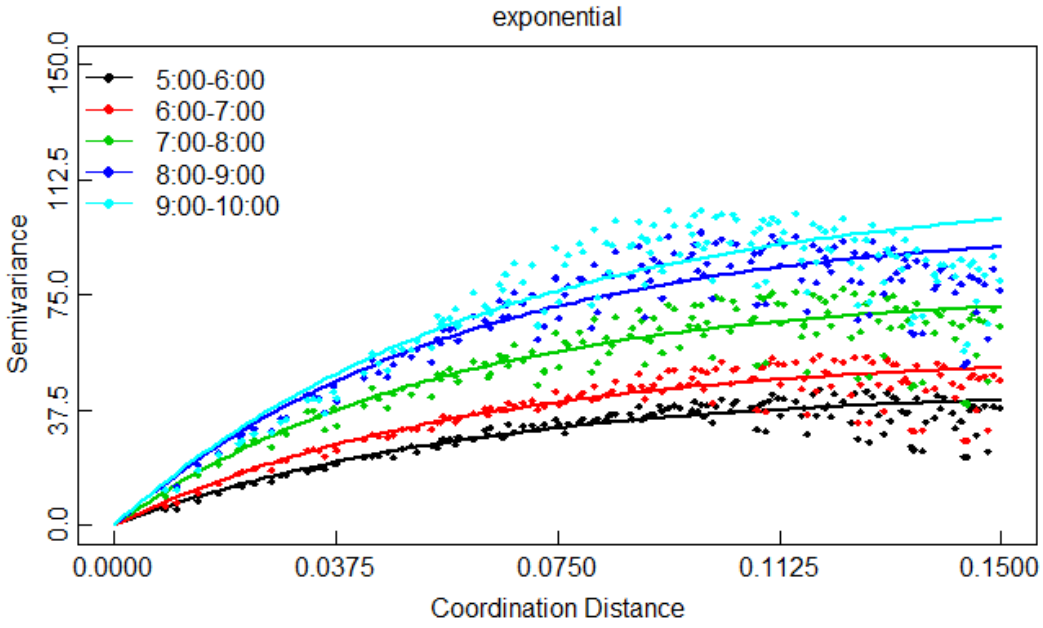

Figure 35: Empirical Fitted Model with Exponential Covariance

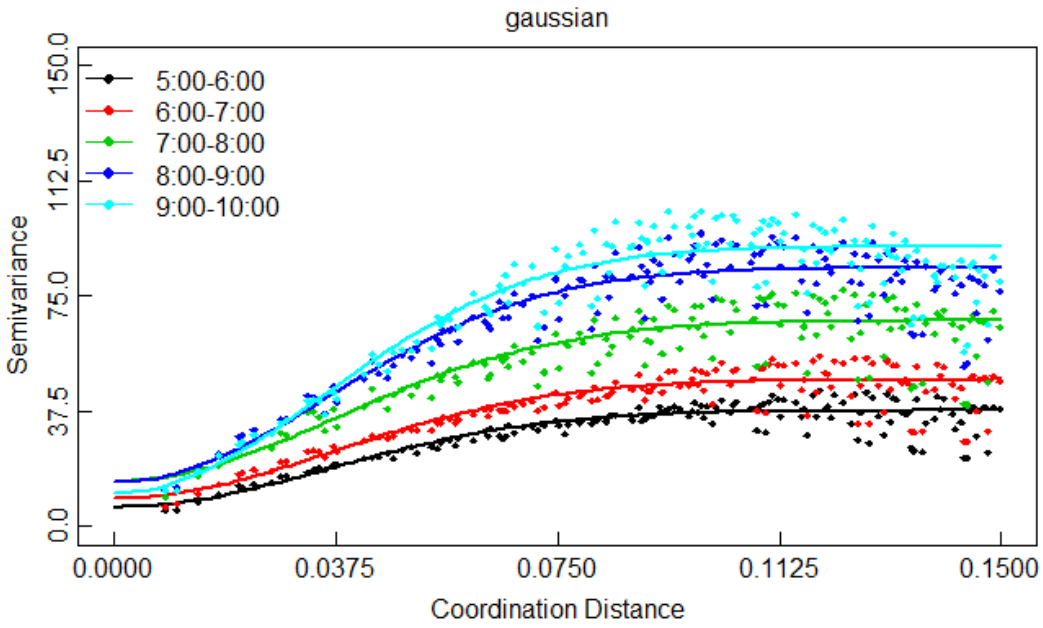

Figure 36: Empirical Fitted Model with Gaussian Covariance

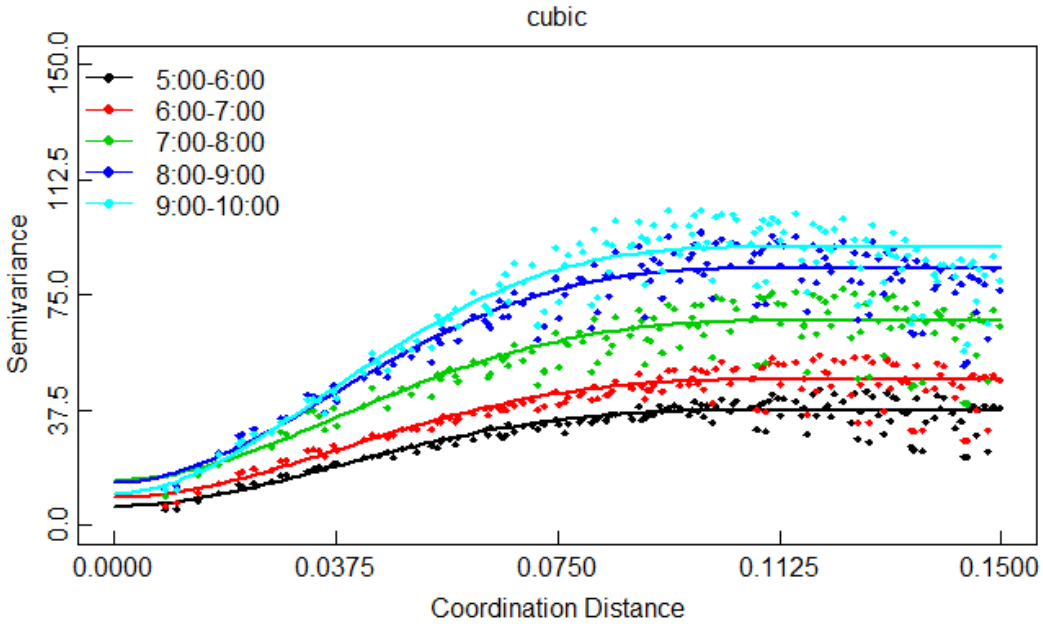

Figure 37: Empirical Fitted Model with Cubic Covariance

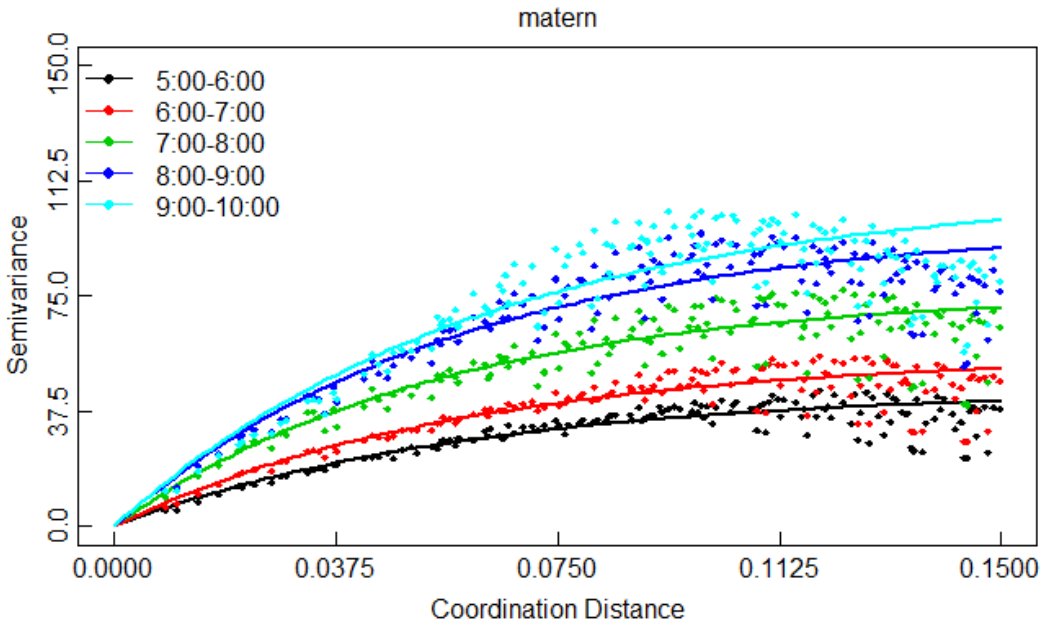

Figure 38: Empirical Fitted Model with Matern Covariance

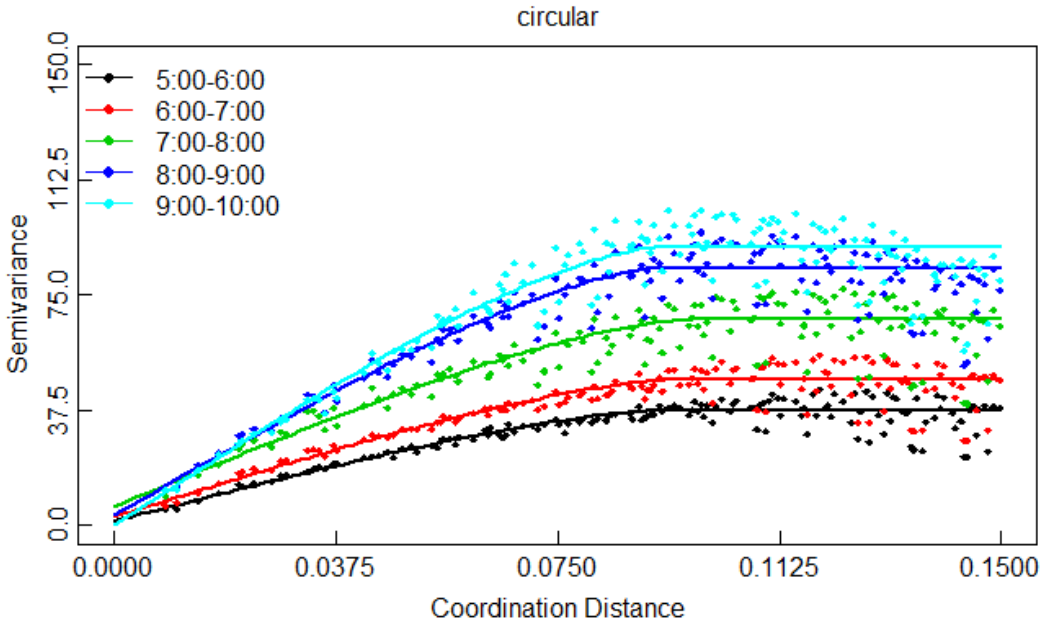

Figure 39: Empirical Fitted Model with Circular Covariance

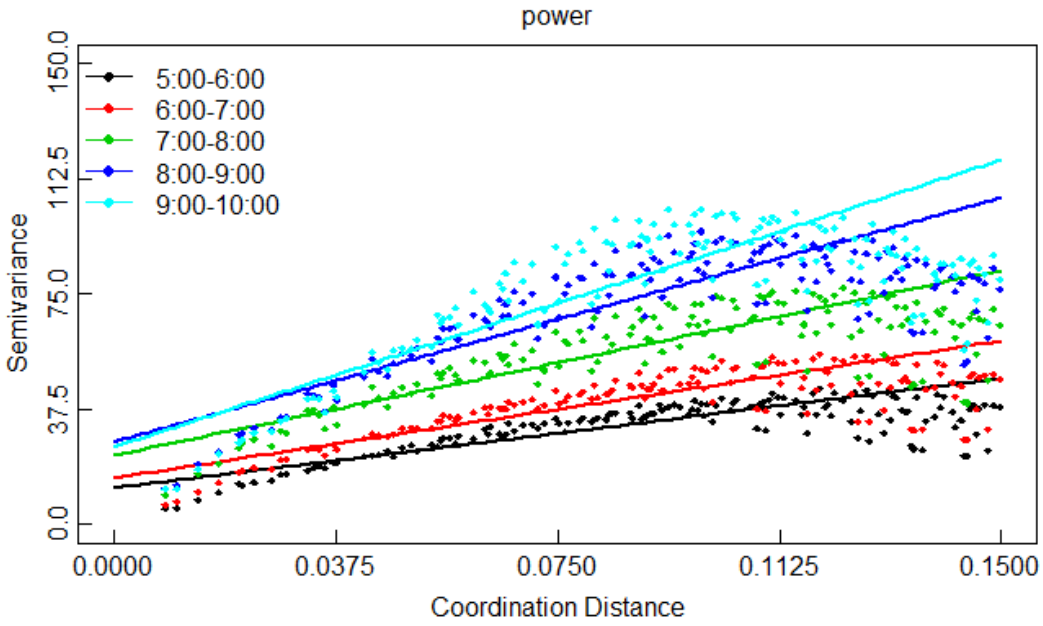

Figure 40: Empirical Fitted Model with Power Covariance

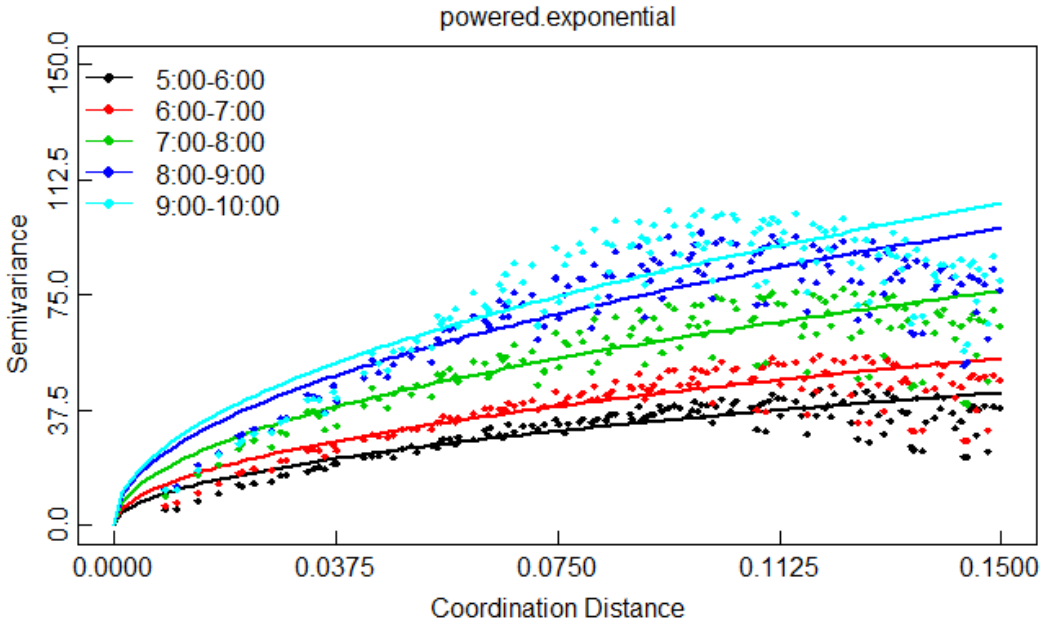

Figure 41: Empirical Fitted Model with Power Exponential Covariance

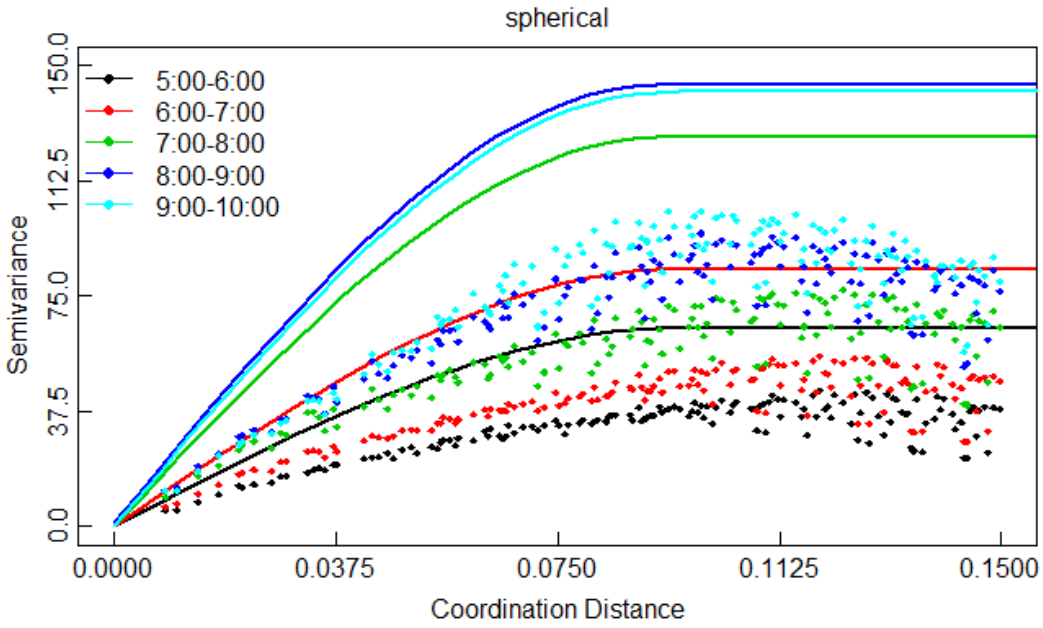

Figure 42: Likelihood Fitted Model with Spherical Covariance

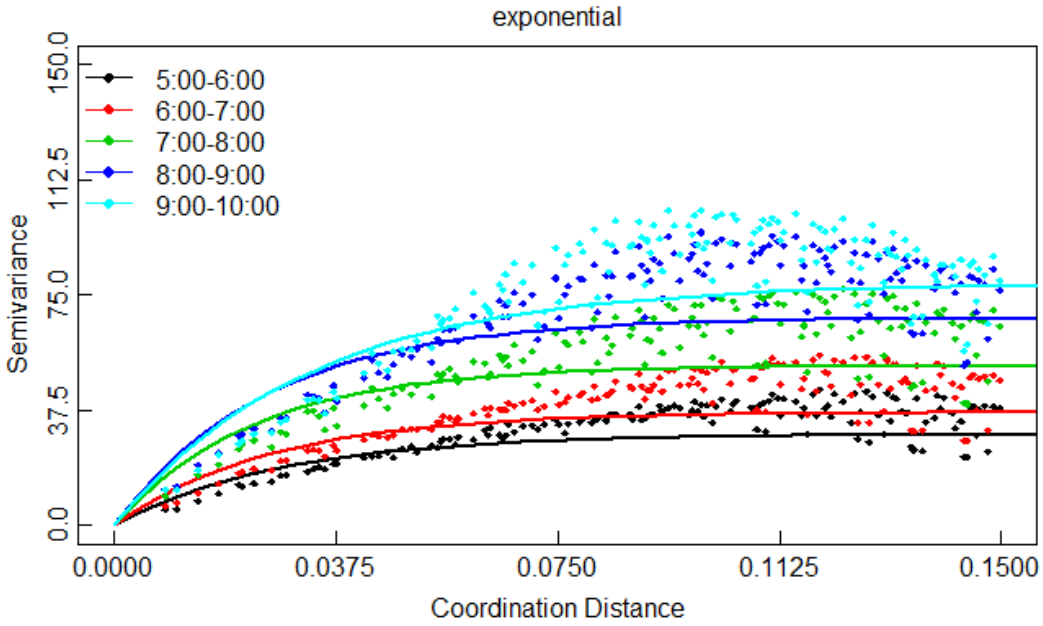

Figure 43: Likelihood Fitted Model with Exponential Covariance

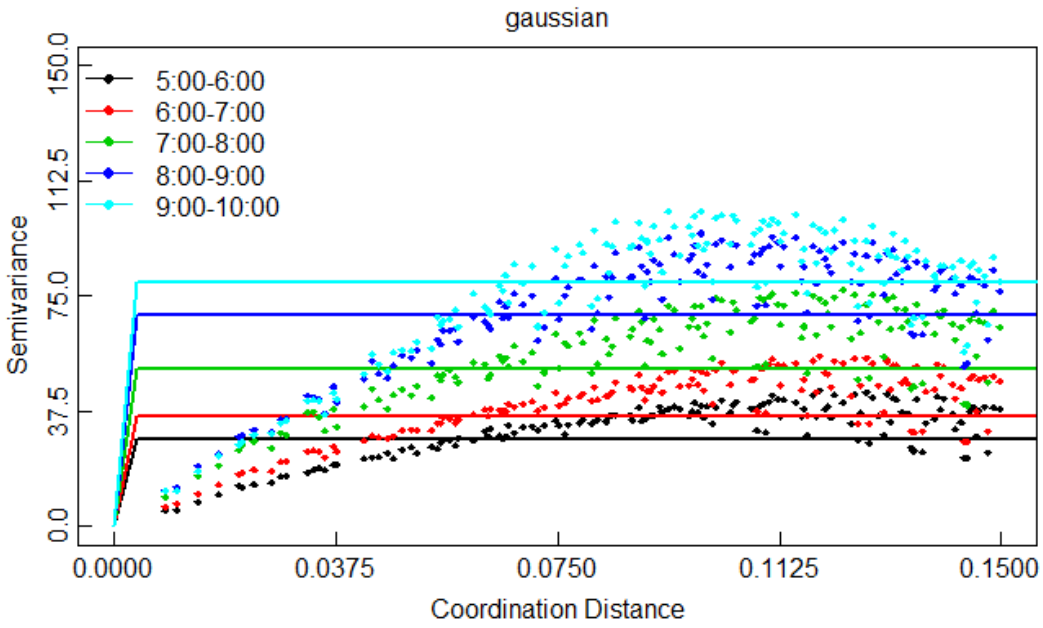

Figure 44: Likelihood Fitted Model with Gaussian Covariance

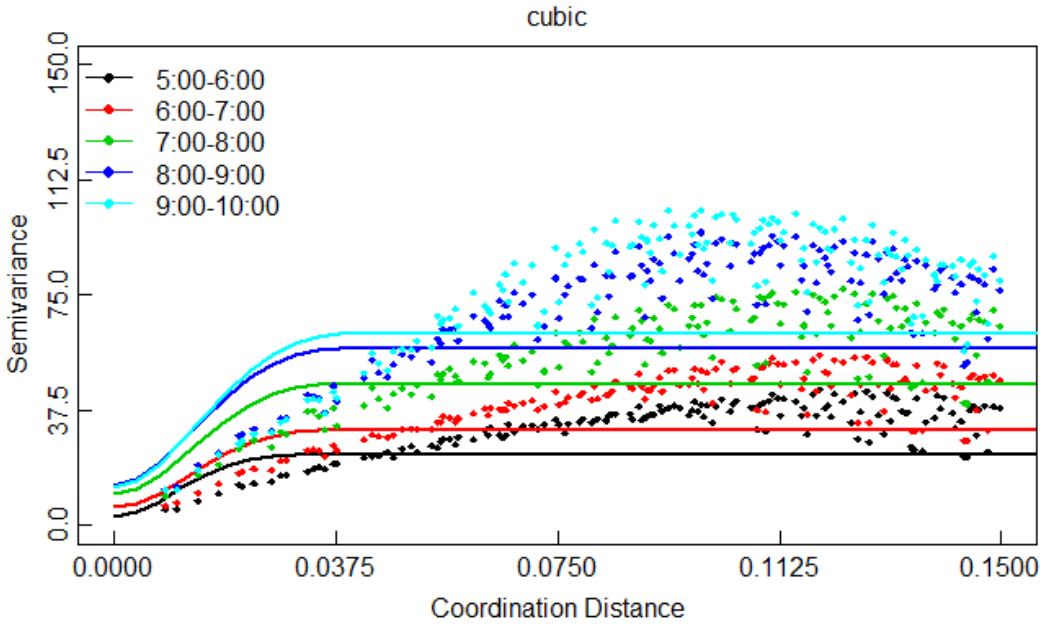

Figure 45: Likelihood Fitted Model with Cubic Covariance

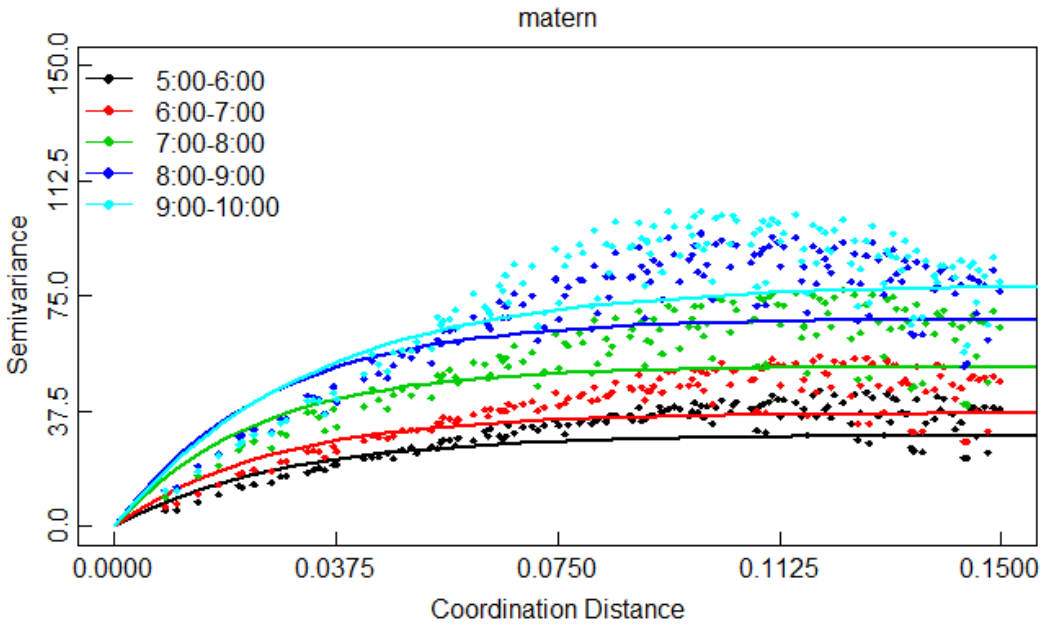

Figure 46: Likelihood Fitted Model with Matern Covariance

B Appendices – Tables

Table 5: Model Fitting for Trend Analysis at 5:00-6:00 am

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4524276582.8114	424064294.7508	-10.67	0.0000
long	69322413.3717	6759171.6736	10.26	0.0000
lat	20175470.1376	4428300.0819	4.56	0.0000
I(long^2)	-301768.0425	27548.8416	-10.95	0.0000
I(lat^2)	-569845.3908	40615.1361	-14.03	0.0000
long:lat	127051.1352	29861.1852	4.25	0.0000

Table 6: Model Fitting for Trend Analysis at 6:00-7:00 am

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5407510455.9421	473626761.5300	-11.42	0.0000
long	83737786.8240	7549149.1031	11.09	0.0000
lat	20687688.3375	4945857.1562	4.18	0.0000
I(long^2)	-365755.1095	30768.6093	-11.89	0.0000
I(lat^2)	-648193.1651	45362.0255	-14.29	0.0000
long:lat	163098.1812	33351.2079	4.89	0.0000

Table 7: Model Fitting for Trend Analysis at 7:00-8:00 am

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6797792159.1983	568114946.2249	-11.97	0.0000
long	106613345.8914	9055198.7031	11.77	0.0000
lat	20773635.1185	5932551.9597	3.50	0.0005
I(long^2)	-468492.8066	36906.9239	-12.69	0.0000
I(lat^2)	-779116.2191	54411.7157	-14.32	0.0000
long:lat	229656.4017	40004.7489	5.74	0.0000

Table 8: Model Fitting for Trend Analysis at 8:00-9:00 am

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7808786493.0229	658937904.9085	-11.85	0.0000
long	122155012.8002	10502828.1716	11.63	0.0000
lat	25071716.4924	6880972.5657	3.64	0.0003
I(long^2)	-539521.6274	42807.1313	-12.60	0.0000
I(lat^2)	-954608.3493	63110.3656	-15.13	0.0000
long:lat	284515.5742	46400.1970	6.13	0.0000

Table 9: Model Fitting for Trend Analysis at 9:00-10:00 am

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7974065064.3337	707087303.4041	-11.28	0.0000
long	124648083.4399	11270282.6696	11.06	0.0000
lat	25962272.8934	7383773.6455	3.52	0.0005
I(long^2)	-551936.4121	45935.1007	-12.02	0.0000
I(lat^2)	-1001344.2093	67721.9172	-14.79	0.0000
long:lat	301227.1079	49790.7161	6.05	0.0000

Table 10: Semivariogram Fitting with Spherical Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	0.80	36.98	0.11	0.11	1345.83
6	1.95	45.93	0.11	0.11	1492.51
7	4.72	62.39	0.11	0.11	1584.26
8	1.96	82.11	0.10	0.10	1087.85
9	0.00	90.89	0.10	0.10	1313.33

Table 11: Semivariogram Fitting with Exponential Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	0.00	44.45	0.06	0.18	1784.89
6	0.00	55.70	0.06	0.18	1777.50
7	0.00	76.33	0.06	0.16	1737.48
8	0.00	97.93	0.06	0.17	1778.25
9	0.00	109.76	0.06	0.19	2688.83

Table 12: Semivariogram Fitting with Gaussian Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	6.28	31.44	0.05	0.09	1585.49
6	9.10	38.71	0.05	0.09	1768.07
7	14.73	52.33	0.05	0.09	1850.30
8	14.30	69.93	0.05	0.09	1292.08
9	10.85	80.29	0.05	0.09	1369.62

Table 13: Semivariogram Fitting with Cubic Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	6.19	31.38	0.12	0.12	1564.83
6	9.03	38.60	0.13	0.13	1744.11
7	14.67	52.14	0.13	0.13	1826.27
8	13.96	69.87	0.12	0.12	1236.83
9	10.35	80.26	0.12	0.12	1278.88

Table 14: Semivariogram Fitting with Matern Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	0.00	44.45	0.06	0.18	1784.89
6	0.00	55.70	0.06	0.18	1777.50
7	0.00	76.33	0.06	0.16	1737.48
8	0.00	97.93	0.06	0.17	1778.25
9	0.00	109.76	0.06	0.19	2688.83

Table 15: Semivariogram Fitting with Circular Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	1.41	36.27	0.09	0.09	1388.80
6	2.84	44.91	0.10	0.10	1544.08
7	6.13	60.85	0.10	0.10	1635.96
8	3.27	80.60	0.09	0.09	1082.57
9	0.00	90.52	0.09	0.09	1200.92

Table 16: Semivariogram Fitting with Power Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	11.80	238.28	1.00	Inf	4369.29
6	15.11	297.83	1.00	Inf	4148.12
7	22.54	401.87	1.00	Inf	3914.71
8	26.87	531.49	1.00	Inf	4643.68
9	25.26	623.86	1.00	Inf	6112.16

Table 17: Semivariogram Fitting with Power Exponential Covariance

	nugget	sigma.sq	phi	range	sum.of.sq
5	0.00	1136.84	100.29	900.02	3177.10
6	0.00	1314.45	83.80	752.10	3017.07
7	0.00	1789.32	78.89	707.96	2784.19
8	0.00	2855.52	125.97	1130.48	3302.95
9	0.00	2885.65	109.40	981.77	4695.43

Table 18: Likelihood Fitting with Spherical Covariance

	nugget	sigma.sq	phi	range	AIC	BIC
5	0.00	64.72	0.10	0.10	4132.69	4151.35
6	0.00	83.96	0.10	0.10	4342.55	4361.21
7	0.00	126.83	0.09	0.09	4689.47	4708.13
8	1.01	143.01	0.09	0.09	4830.25	4848.91
9	0.00	141.49	0.09	0.09	4771.84	4790.50

Table 19: Likelihood Fitting with Exponential Covariance

	nugget	sigma.sq	phi	range	AIC	BIC
5	0.00	29.72	0.03	0.08	4108.04	4126.70
6	0.00	36.85	0.03	0.08	4315.91	4334.56
7	0.00	52.04	0.02	0.07	4659.61	4678.27
8	0.00	67.68	0.03	0.08	4805.03	4823.69
9	0.00	78.78	0.03	0.10	4756.67	4775.33

Table 20: Likelihood Fitting with Gaussian Covariance

	nugget	sigma.sq	phi	range	AIC	BIC
5	0.00	28.52	0.00	0.00	4859.87	4878.53
6	0.00	35.58	0.00	0.00	5033.19	5051.85
7	0.00	51.19	0.00	0.00	5318.42	5337.07
8	0.00	68.87	0.00	0.00	5550.96	5569.62
9	0.00	79.30	0.00	0.00	5661.54	5680.20

Table 21: Likelihood Fitting with Cubic Covariance

	nugget	sigma.sq	phi	range	AIC	BIC
5	2.80	20.05	0.03	0.03	4123.60	4142.26
6	5.78	25.24	0.04	0.04	4336.64	4355.30
7	10.16	35.98	0.04	0.04	4684.59	4703.25
8	13.12	44.67	0.04	0.04	4854.45	4873.10
9	12.18	50.48	0.05	0.05	4831.22	4849.88

Table 22: Likelihood Fitting with Matern Covariance

	nugget	sigma.sq	phi	range	AIC	BIC
5	0.00	29.72	0.03	0.08	4108.04	4126.70
6	0.00	36.85	0.03	0.08	4315.91	4334.56
7	0.00	52.04	0.02	0.07	4659.61	4678.27
8	0.00	67.68	0.03	0.08	4805.03	4823.69
9	0.00	78.78	0.03	0.10	4756.67	4775.33

Table 23: Gamma GLM with No Spatial Effect

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0000	15.52	0.0000
D2Metro	0.0000	0.0000	4.82	0.0000
D2Road	-0.0000	0.0000	-1.22	0.2236
I(D2Metro^2)	0.0000	0.0000	10.23	0.0000
I(D2Road^2)	0.0000	0.0000	3.08	0.0021
D2Metro:D2Road	0.0000	0.0000	1.77	0.0768

Table 24: Negative Binomial GLM with No Spatial Effect

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.6650	0.0412	234.76	0.0000
D2Metro	-0.0006	0.0000	-15.83	0.0000
D2Road	0.0004	0.0001	8.55	0.0000
I(D2Metro^2)	0.0000	0.0000	4.39	0.0000
I(D2Road^2)	-0.0000	0.0000	-7.79	0.0000
D2Metro:D2Road	-0.0000	0.0000	-16.32	0.0000

Table 25: Negative Binomial GAM with Spatial Effect

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.95	0.04	244.15	0.00
D2Metro	-0.00	0.00	-8.68	0.00
D2Road	0.00	0.00	7.33	0.00
I(D2Metro^2)	0.00	0.00	2.47	0.01
I(D2Road^2)	-0.00	0.00	-6.79	0.00
D2Metro:D2Road	-0.00	0.00	-7.83	0.00

C Appendices – R codes