STAT 5544 Final Report

Spatial Analysis

Fall 2016

**Spatial Pattern of Population Movement Using Cellular Data**

Wenyu Gao, Danni Lu

# Abstract

As increasing number of people are living in urban area, rising issue 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 origin and how population moves spatially. In this report, we proposed a method to analysis spatiotemporal 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 semivariogram structure were fitted and compared. Predictions were made based on fitted semivariogram structure. Prediction result was displayed in choropleth map and the population movement pattern is discussed. The result showed that during the morning peak hours, population move from all over the city to central areas. Areas with relative high population density expand, grow and converging along transportation corridors.

# Keywords:

Periodical Traffic, Population Movement, Spatial Analysis, Cellular data.

# Introduction

Recurrent traffic congestion in central business districts is a major sources of cost for transportation system and environments such as pollution. Worldwide, urban population has been growing smoothly during the past decades, from 33.56% of the world’s population in 1960 to 53.86% in 2015. In developed country, the proportion is even higher. 82% of total population in United States was from urban area in 2015([World](http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS) Bank, 2015). As increasing people are living in urban area, the rising issues such as limited land resource and deteriorating periodical traffic congestion are gaining more and more attention. One of the major causes of periodical traffic congestion is overdevelopment in districts where the traffic attraction exceeds the accommodation volume of traffic system. Central Business Districts (CBD) are therefore becoming controversial as they’re key areas that have been bringing tremendous pressure to transportation system during peak hours in weekday due to their high traffic attraction (Willett K, 2006).

The CBDs is a major destination for ubran area and is major source of traffic demand. Central Business District (CBD) is the commercial and business center of a city with high developing density. The 21st century CBD within metropolitan areas usually characterized by a concentration of commercial and business buildings, along with other public infrastructure such as entertainment, shopping malls, medical centers (Olayiwola K O, 2014). According to a traffic survey of 63 cities in US in 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 of great convenience 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 in a typical weekday (Mindali O, 2004). Especially in big metropolitan like New York, Sydney and Shanghai. The fact that CBDs are usually geographical center of a city makes the traffic situation even worse. Even though CBDs are facilitated with mass transit system and high capacity road system, the tremendous traffic they attract are usually difficult to accommodate. As a consequence, the traffic congestion in the CBD areas has become a bottleneck for a better city life for every growing metropolitan.

Alleviating periodical traffic congestion is a long term strategy for metropolitan. The very first step should be figuring out the spatial and temporal characteristic of travel behavior during peak hours. This is a classical topic in transportation planning and a lot work has been done since decades ago. Most of the related inferences have been made from traffic survey data. For example, establishing models to compute travel time, traffic volume for roads and evaluate optimal routes based on OD survey data (Thériault, 1999). In early 1950s, Foley and his team collected traffic survey data from municipal officials from more than fifty cities to explore characteristics of travel related to CBD and summarized the quantity and ratio of population who has daily trip destinations in CBD in weekdays (Foley D L, 1954).

Traditional method to acquire the travel pattern information is conducting household survey, which is usually inefficient and expensive. Nowadays, high density of cellular tower and deep penetration of mobile phones provides us with detailed travel information with high accuracy and precise timing. Some previous study has shown a great potential of mobile data in exploring commuting traffic. As early as 20 years ago, Maryland State Highway Administration conducted a prior project to assess the viability of using cellular-based traffic probes as traffic surveillance technique and demonstrated the technical potential to provide information like travel location, speed and traffic conditions (CAPITAL, 1997). Later, simulation studies were conducted by Virginia Department of Transportation to test the performance of Probe-based traffic monitoring systems. They demonstrated that sampling based on cellular coverage areas is better than other sampling approach in estimating road speed in terms of availability and accuracy (Fontaine, 2007). It’s not until recently that researchers begin to explore travel pattern using real cellular dataset of large scale. Instead of zonal sampling, real cellular dataset has wider coverage and is better representative of population, which allows researchers to study all kinds of travel patterns instead of calculating driving speed of certain road. Cellular data provide us a new way of looking at cities as a dynamic system with detailed information of urban mobility and travel behavior (Reades et al. 2007). Researchers in Beijing used cellular data to calculate population inflow and outflow volume of selected traffic zone (Dong, Honghui, 2014). Researchers from MIT showed the relationship between home-work travel time and travel distance based on mobile phone call detail records (CDRs). Travel time in Minnesota was also estimated using Cell phone data from Sprint PCS Mobile network (Liu, H. X, 2008).

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 for population in a whole city area. Midtown in Shanghai is one of the largest central business districts in the world, which makes it a perfect study area for 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. First, we established a spatial model to make inference of population density of whole study area. Spatial pattern of population density during each hour in the morning was displayed and the temporal trend of population density was discussed. On the basis, we proposed a method to identify Top Traffic Attraction Zones that attracted most traffic during morning peak hours in a typical weekday. People gathered in Top Traffic Attraction Zones during morning peak are targeted and their trip origins are extracted. At last, we discussed the characteristic of the targeted trips, including the spatial pattern of trip origins, arriving time and distribution of trip distance.

# 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 in 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).

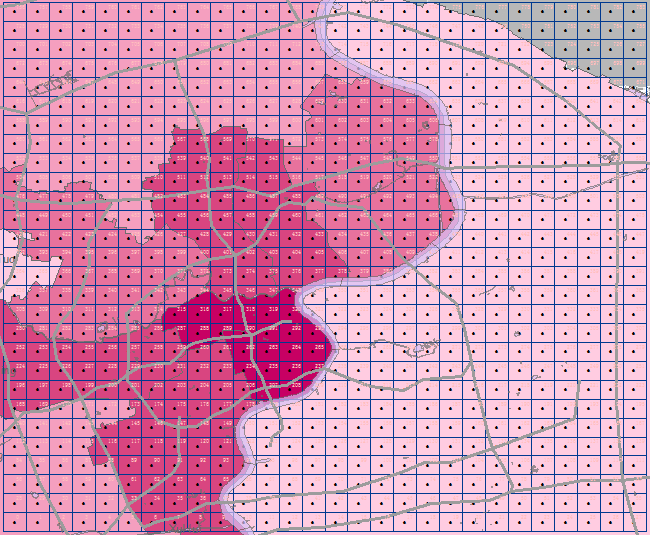
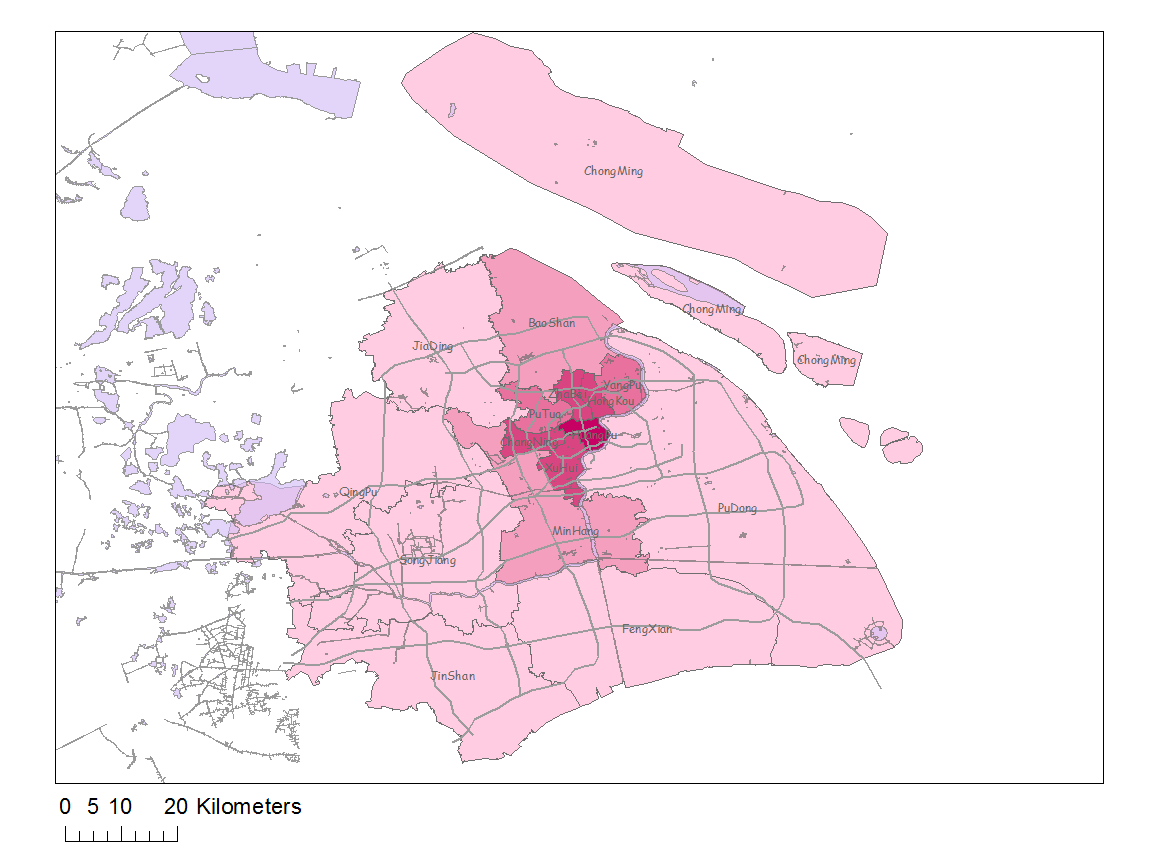
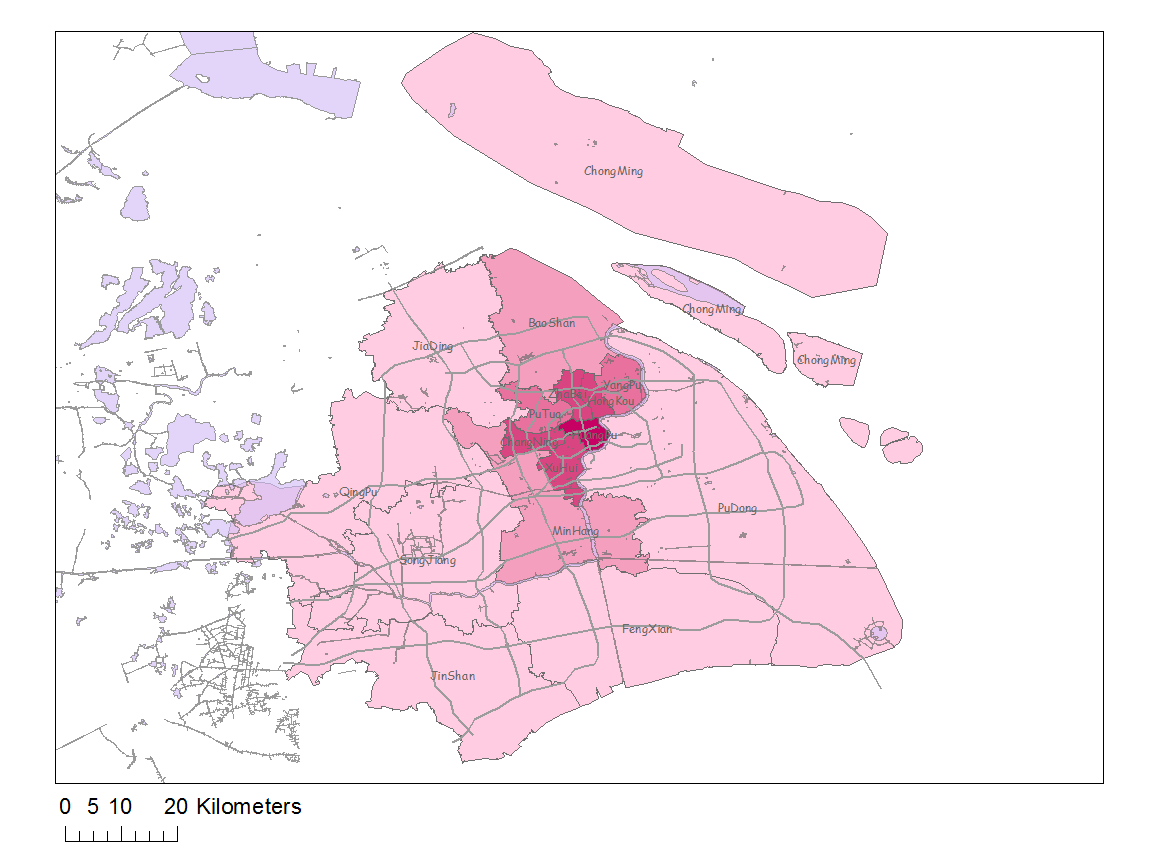


Figure 1. Study Area and Aggregation Grid

The accuracy of location information we get depends on the density of cell towers. Instead of accurate location of mobile phone users, cellular data recorded 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. Previous study shows 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.

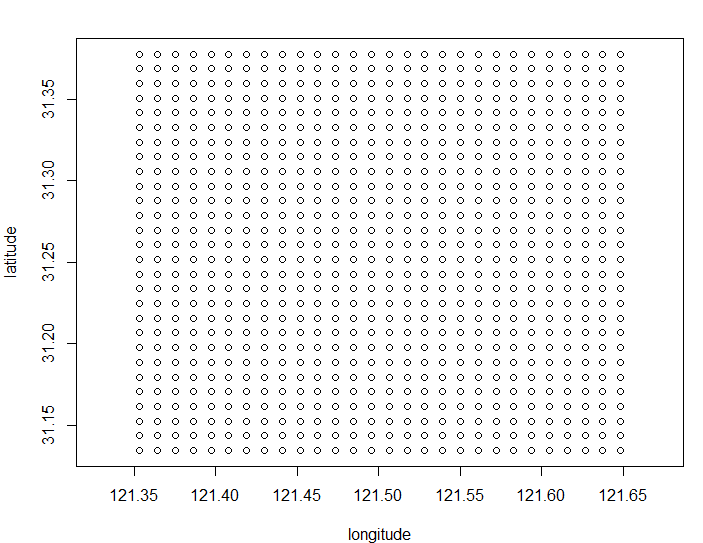
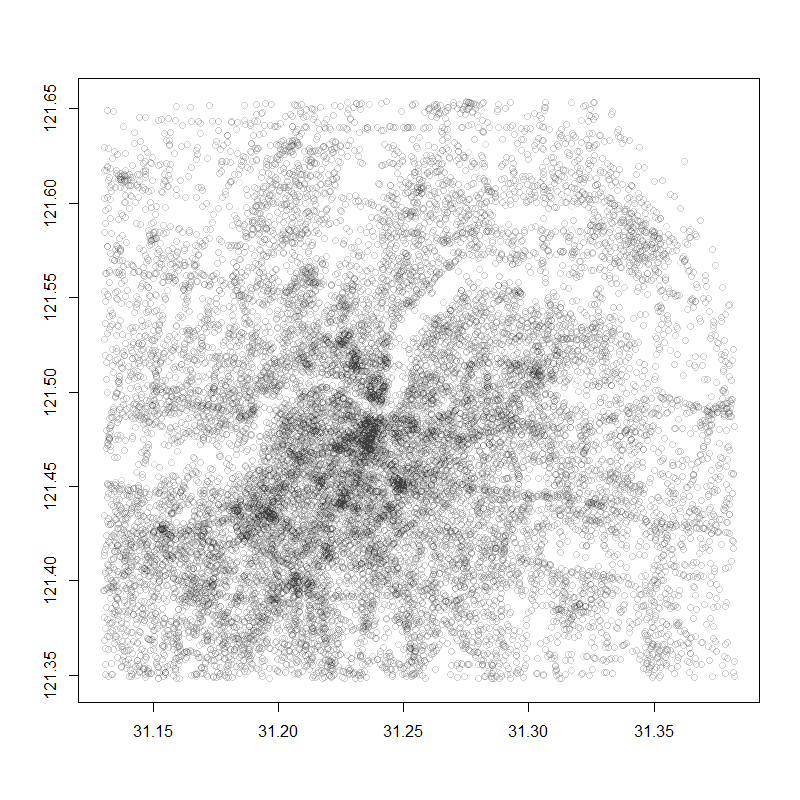
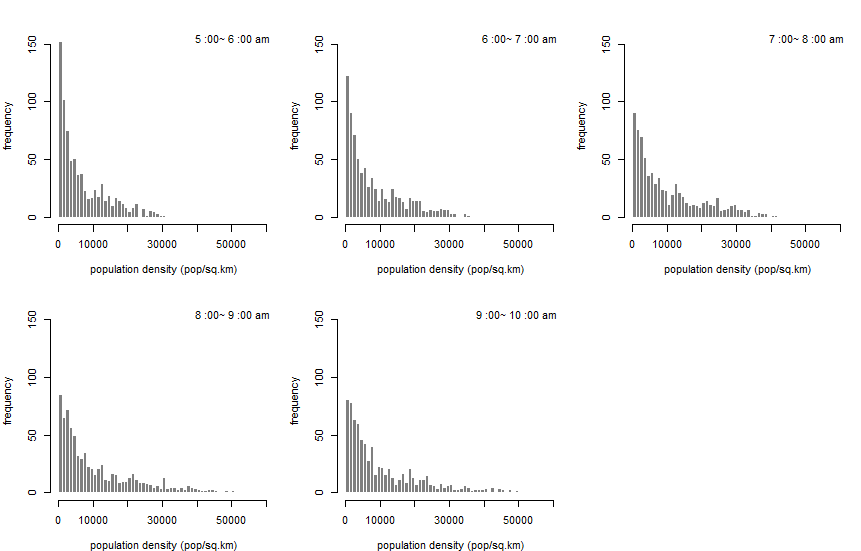


Figure 2. Original Cell Towers and Aggregated Cell Towers

# Exploratory Data Analysis

**Distribution of response and covariates**



Histogram of population density in each hour

Population Density Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| **5:00-6:00** | 0 | 1378 | 4217 | 7003 | 11200 | 33550 |
| **6:00-7:00** | 0 | 1849 | 5322 | 8421 | 13430 | 36390 |
| **7:00-8:00** | 0 | 2452 | 6940 | 10550 | 16050 | 47720 |
| **8:00-9:00** | 0 | 2588 | 7033 | 11440 | 16990 | 68800 |
| **9:00-10:00** | 0 | 2554 | 6772 | 11450 | 17050 | 83820 |

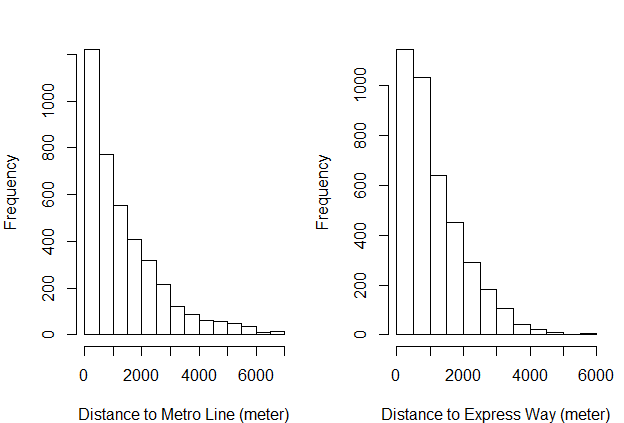
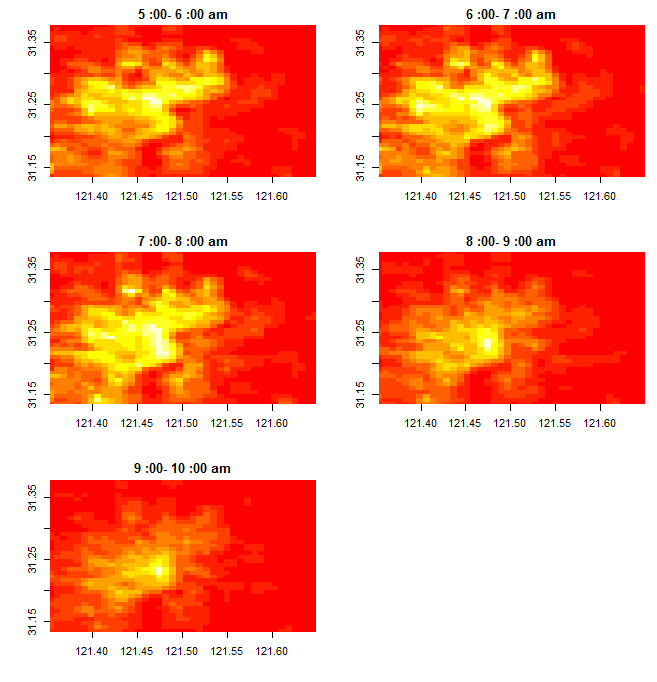
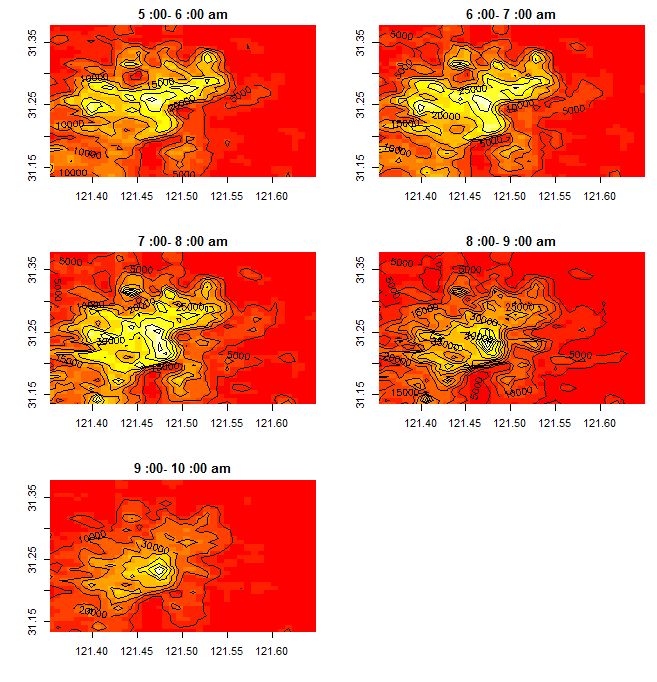


Figure. Histogram of covariates

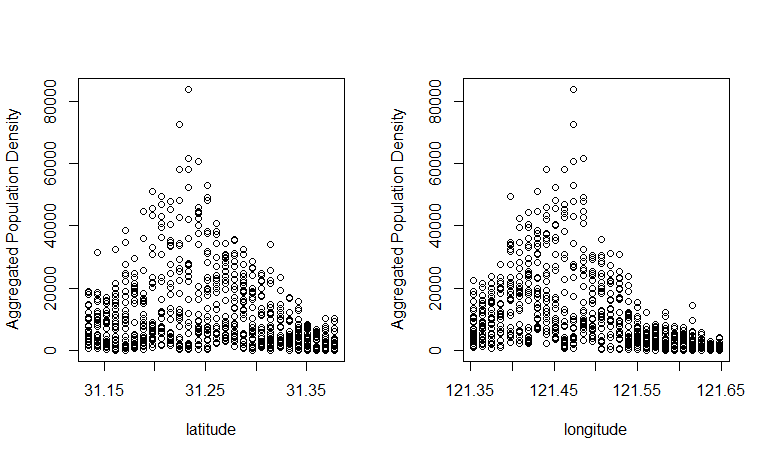
**Illustration spatial plot**



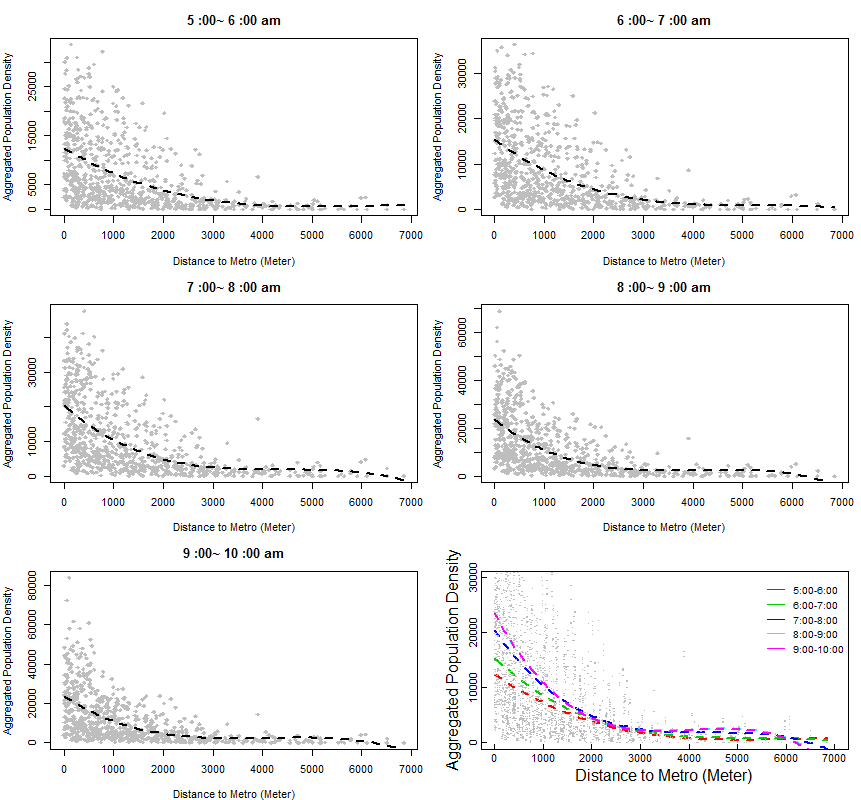


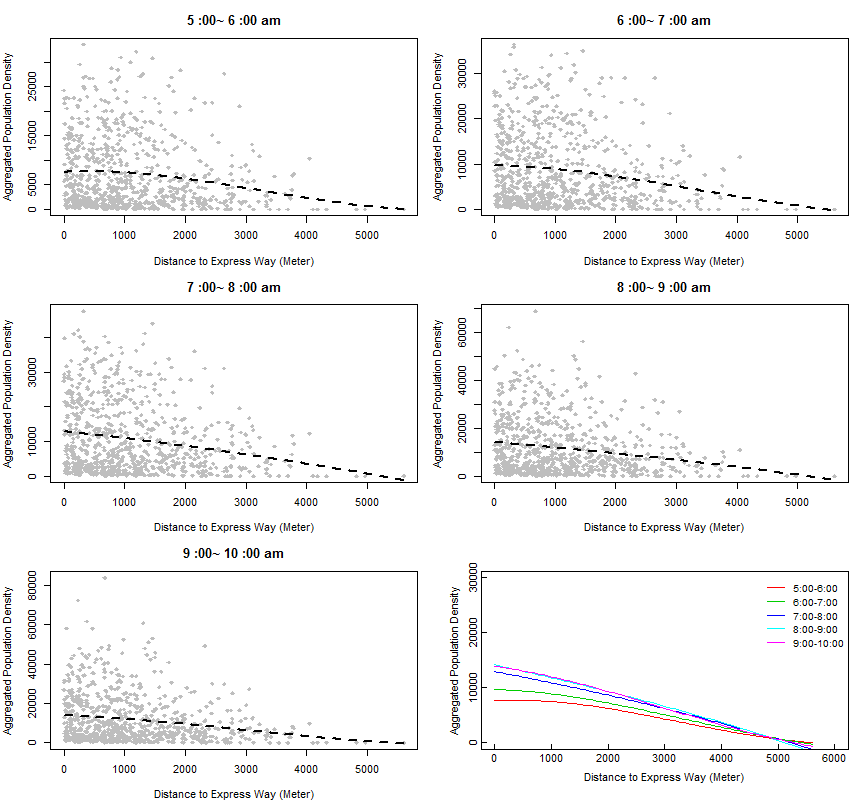
# Spatial Trend and Semivariogram

**Regression on Coordination**

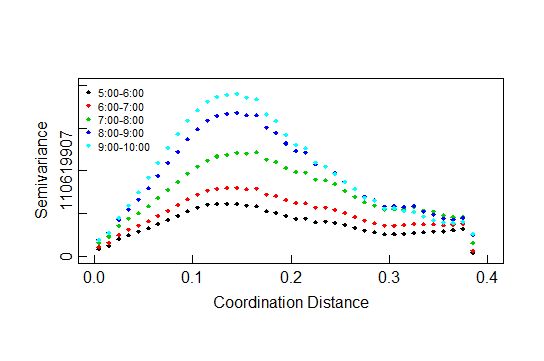


**Regression on Covariates**

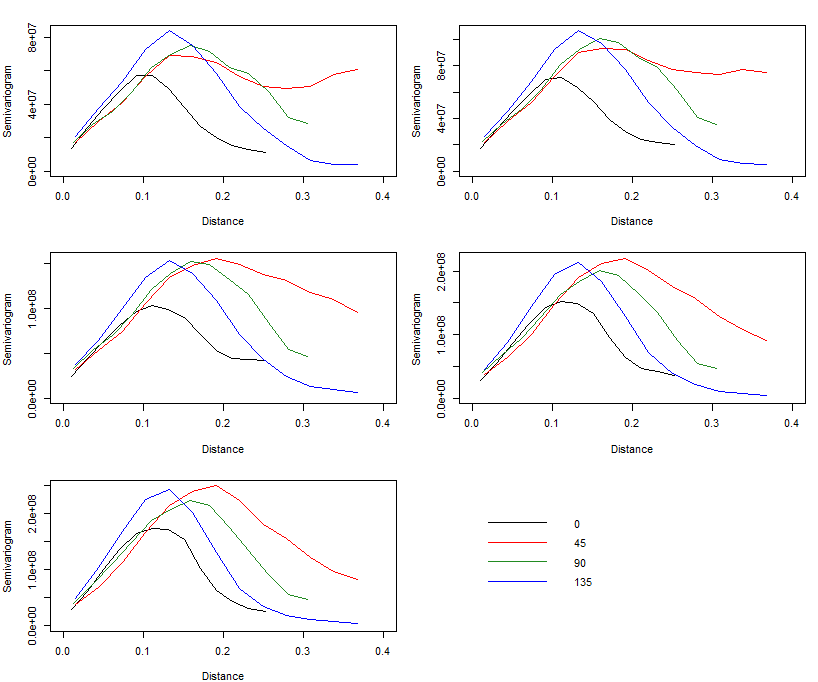




**Semivariogram**

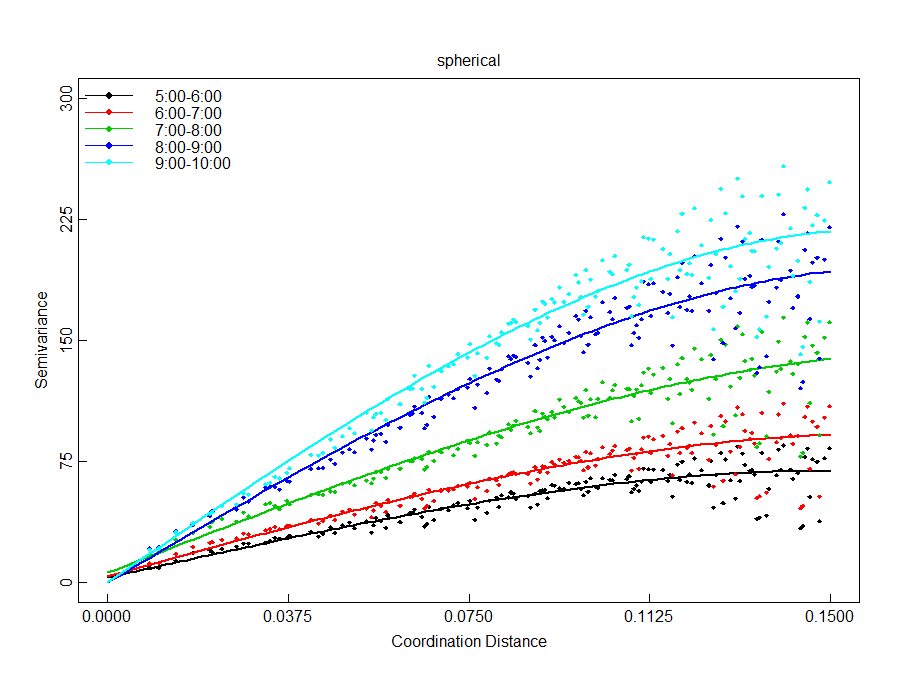
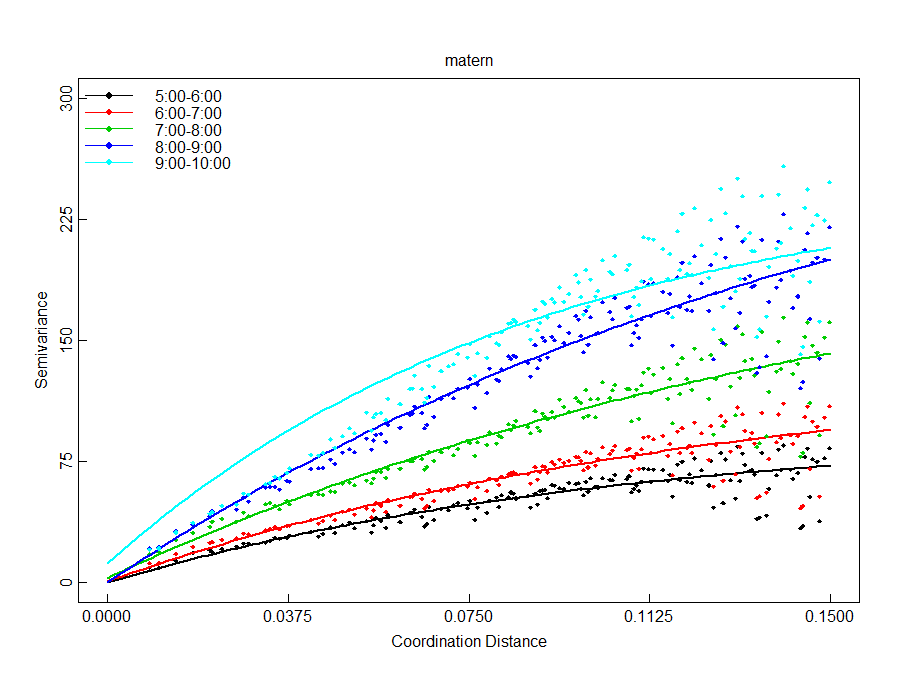
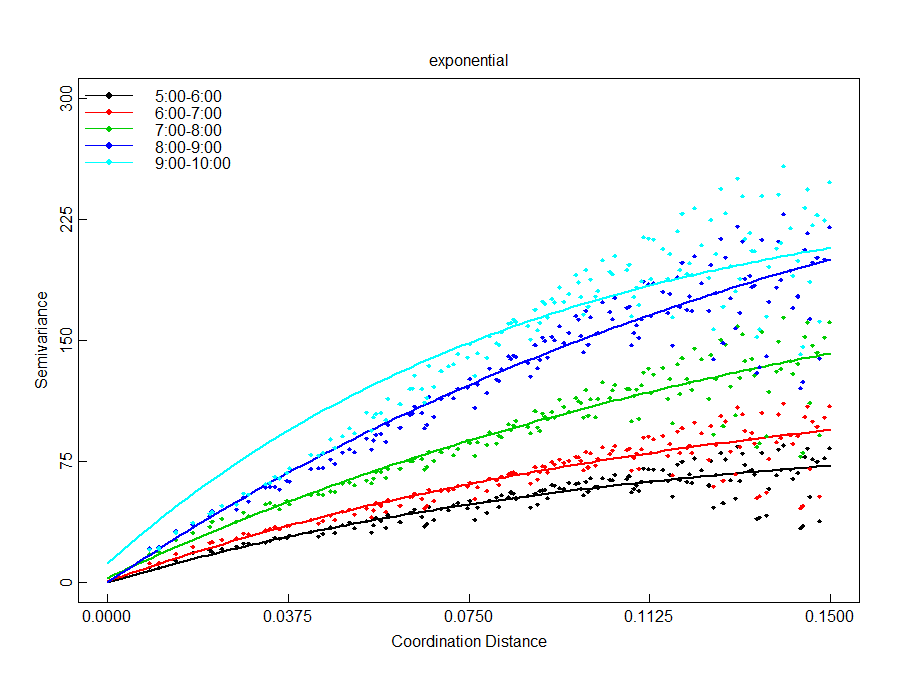
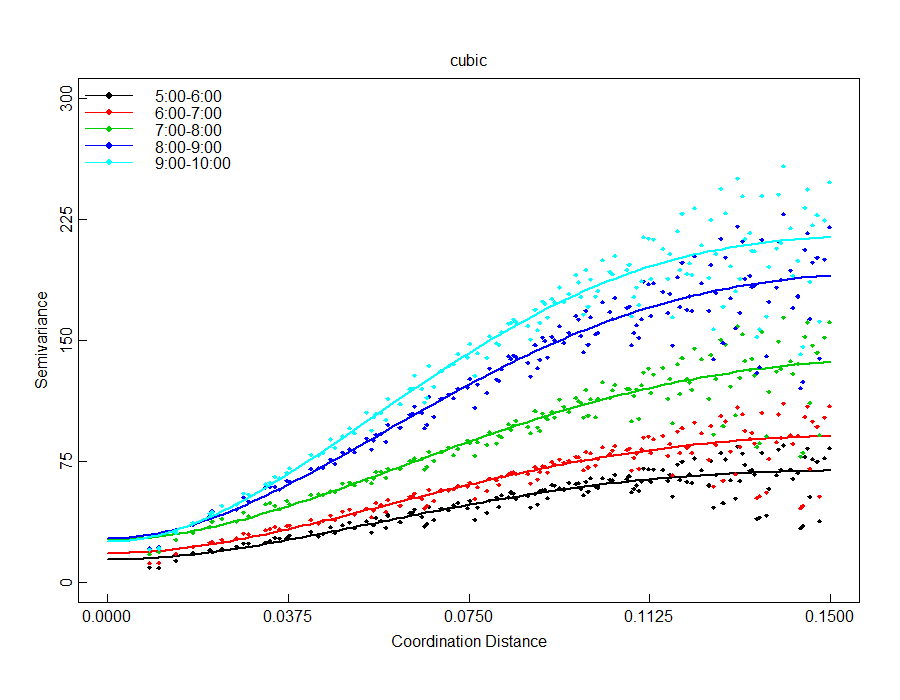
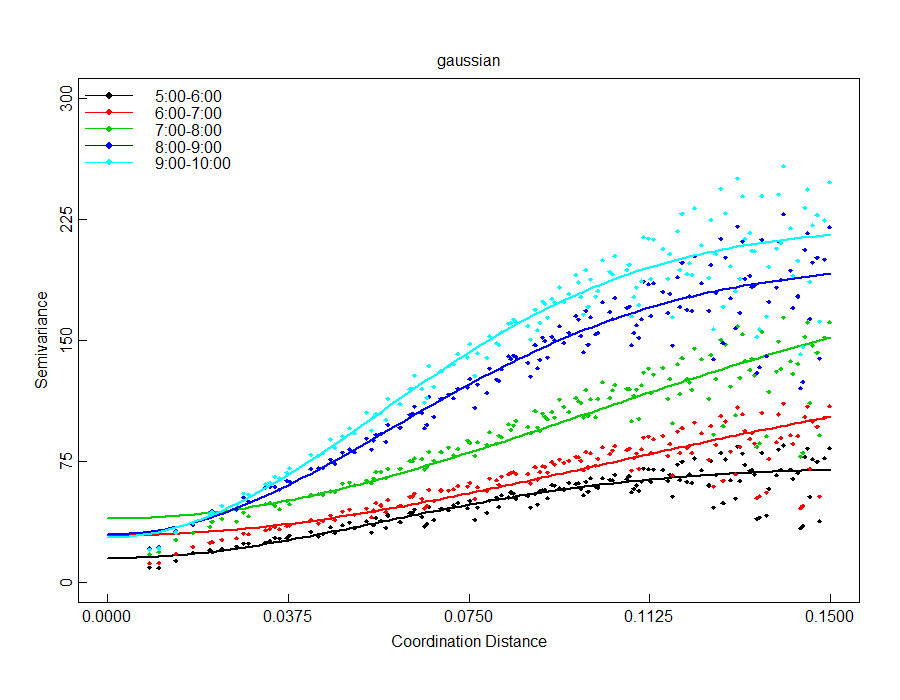


**Anisotropy**



# Model Fitting and Prediction

**Semivariogram Fitting**



**Model Compare**

MSE

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **nugget** | **sigma.sq** | **phi** | **range** | **sum.of.sq** | **total sum** |
| **Gaussian** |  |  |  |  |  |  |
| 5:00-6:00 | 15.0781 | 56.2032 | 0.0793 | 0.1372 | 7976783 | 120239822 |
| 6:00-7:00 | 29.0973 | 112.4276 | 0.1456 | 0.2519 | 16653292 |
| 7:00-8:00 | 39.5578 | 165.0949 | 0.1405 | 0.2432 | 27832245 |
| 8:00-9:00 | 29.6936 | 167.9432 | 0.0835 | 0.1445 | 30554602 |
| 9:00-10:00 | 27.8238 | 192.3736 | 0.0788 | 0.1363 | 37222900 |
| **Cubic** |  |  |  |  |  |  |
| 5:00-6:00 | 14.2211 | 55.4374 | 0.1826 | 0.1826 | 7924832 | 108493818 |
| 6:00-7:00 | 18.1704 | 73.799 | 0.1924 | 0.1924 | 12204207 |
| 7:00-8:00 | 26.5823 | 113.2998 | 0.2037 | 0.2037 | 20961528 |
| 8:00-9:00 | 27.2836 | 165.2804 | 0.1922 | 0.1922 | 30410606 |
| 9:00-10:00 | 25.3419 | 189.7097 | 0.1821 | 0.1821 | 36992645 |
|  |  |  |  |  |  |  |
| **Spherical** |  |  |  |  |  |  |
| 5:00-6:00 | 3.0787 | 66.0419 | 0.1496 | 0.1496 | 7956700 | 112843207 |
| 6:00-7:00 | 4.0104 | 87.9961 | 0.1607 | 0.1607 | 12340649 |
| 7:00-8:00 | 6.0968 | 136.7066 | 0.1763 | 0.1763 | 21364196 |
| 8:00-9:00 | 0 | 195.4007 | 0.166 | 0.166 | 31840797 |
| 9:00-10:00 | 0 | 218.7919 | 0.1603 | 0.1603 | 39340865 |
|  |  |  |  |  |  |  |
| **Matern** |  |  |  |  |  |  |
| 5:00-6:00 | 0 | 95.7051 | 0.1065 | 0.3191 | 8344750 | 146254407 |
| 6:00-7:00 | 0.7393 | 136.2881 | 0.1285 | 0.3851 | 12803426 |
| 7:00-8:00 | 2.8467 | 234.2681 | 0.1664 | 0.4984 | 22033151 |
| 8:00-9:00 | 0 | 338.2249 | 0.1671 | 0.5007 | 34633507 |
| 9:00-10:00 | 11.7887 | 240.0527 | 0.0891 | 0.267 | 68439573 |
|  |  |  |  |  |  |  |
| **Exponential** |  |  |  |  |  |  |
| 5:00-6:00 | 0 | 95.7051 | 0.1065 | 0.3191 | 8344750 | 146254407 |
| 6:00-7:00 | 0.7393 | 136.2881 | 0.1285 | 0.3851 | 12803426 |
| 7:00-8:00 | 2.8467 | 234.2681 | 0.1664 | 0.4984 | 22033151 |
| 8:00-9:00 | 0 | 338.2249 | 0.1671 | 0.5007 | 34633507 |
| 9:00-10:00 | 11.7887 | 240.0527 | 0.0891 | 0.267 | 68439573 |

**Prediction: Kriging Models and Natural Neighbor**

To evaluate the spatial pattern of the population density, we used the Kriging method. The method could provide a smooth surface of population density. The first step in Kriging method is fitting semivariogram model, which provides crucial quantitative evaluation on the spatial patterns of population density. We applied the method to the data from 5:00am to 10:00am.

For the aggregated cellular data, we consider the aggregated cell towers in the geo-center of each areal unit as sampled sites. In total, we have 784 evenly spaced sites, denoted as. We then use the information from discrete sampled sites to make inference about the continuous field of study area. Since each areal unit is one square kilometer, number of mobile users in aggregated cell tower provides a reasonable estimate for population density of the site. Assuming there is no difference in ownership among different areas in Shanghai, the number of mobile users of sampled site is proportional to true population density. Population movement is a continuous spatial process. As a result, locations near a high population density site are more likely to be high density. Generally, we assume the spatial association between two sites mainly depends on distance. That is, the correlation between two sites can be expressed as a function of distance,

where *d* is distance between two sites is number of mobile user at site . Semivariogram is an important function to indicate spatial correlation. In isotropy cases, semivariogram is defined as

.

The calculation form for semivariogram is . A larger semivariogram indicating larger variability in the random field, which in normalized case suggesting weaker spatial correlation.

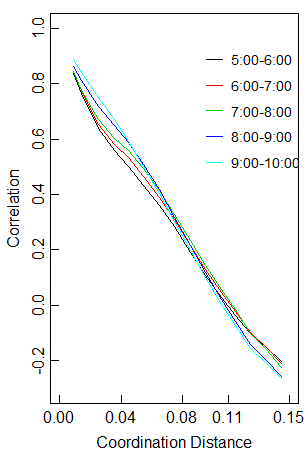
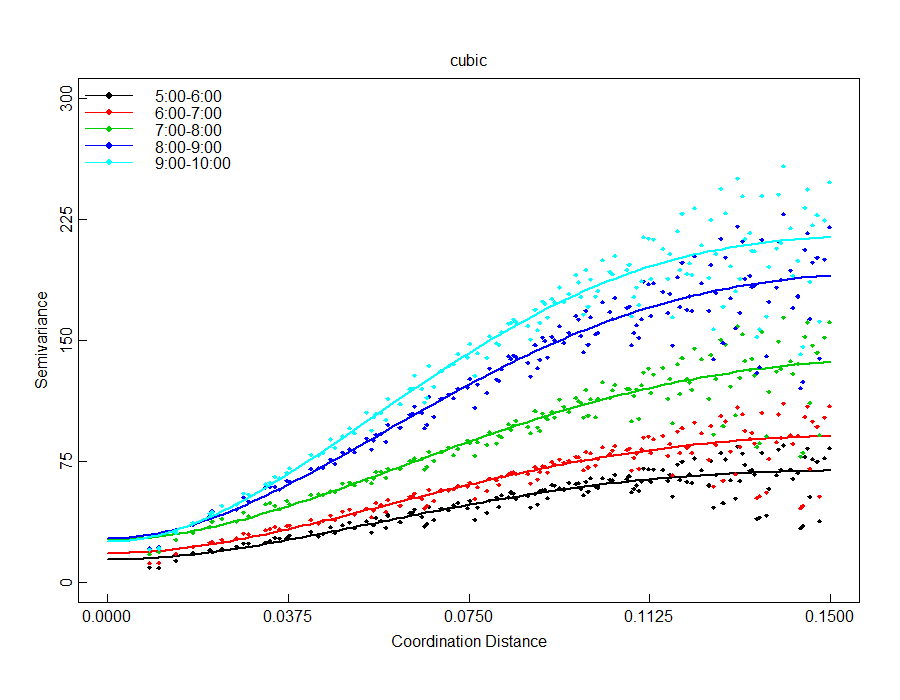
To calculate empirical semivariogram using cellular data within study area, we use Hawkins-Cressie semivariogram estimator (Hawkins and Cressie, 1980) which gives relative robust estimate in case of outliers.

In the study, we treat our study area as a continuous random field. In this field, population movement is considered as a continuous process. As a result, the population density at each particular time should be a smooth surface. During the process of data collecting, data cleaning and data aggregating, we’re unavoidably involving random errors. In order to eliminate the effect caused by random error and focus on the main trend of spatial pattern, we use Ordinary Kriging method to make prediction of study area using aggregated data. Kriging is a continuous interpolating method for spatial data. It weights the surrounding measured values to derive a prediction for an unmeasured location. The general formula for Kriging is formed as a weighted sum of the data:

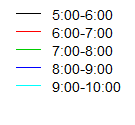
We solve for the estimate that minimize squared error loss (Banerjee, etc. 2014):

Thus, we incorporate the semivariogram structure in the Kriging model to predict unmeasured locations.

We obtained empirical semivariogram using Hawkins and Cressie method mentioned above. We then further fitted five different semivariogram models: Gaussian, Cubic, Spherical, Matern and Exponential. Using least square method, Cubic model turned out to be best fit in terms of minimizing total residual sum of square. Fitted curve for each time period is shown in Figure 4 on top of scatter points of empirical semivariogram. Corresponding estimated model parameters for each time period are shown in Table 1. The plot shows an increasing trend of semivariogram with respect to spatial distance. Theoretically, at zero separation distance, the semivariogram value is 0. However, at an infinitesimally small separation distance, the semivariogram often exhibits a nugget effect, which is some value greater than 0. In our fitted plots, the intercepts for each hour range from 14.22 to 27.28. This can be attributed to measurement errors or variation at distances smaller than the aggregation interval or both. We also notice that at a certain distance, the model levels out. The distance where the model first flattens out is known as the range. Sample locations separated by distances smaller than the range are spatially autocorrelated, whereas locations farther apart than the range are not. According to the fitted model, the range for different time periods are similar. Average speaking, locations within 19 km are correlated while locations further apart are not. The max semivariogram value attained at the range is called the sill, indicating the upper limit of variation across the study region.



0 4 8 12 16 0 4 8 12 16

Distance (Km)  Distance (Km)

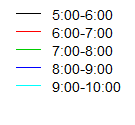
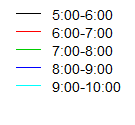


Figure 4. Semivariogram and Fitted Models

Table 1. Estimated Cubic Model Parameters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Nugget** | | **Sill** | | **Phi** | **Range** |
| 5:00 ~6:00 | | 14.2211 | | 69.6585 | 0.1826 | 0.1826 |
| 6:00 ~7:00 | | 18.1704 | | 91.9694 | 0.1924 | 0.1924 |
| 7:00 ~8:00 | | 26.5823 | | 139.8821 | 0.2037 | 0.2037 |
| 8:00 ~9:00 | | 27.2836 | | 192.564 | 0.1922 | 0.1922 |
| 9:00 ~10:00 | | 25.3419 | | 215.0516 | 0.1821 | 0.1821 |

As is shown in Figure 4, the upper limit of variation in population density across study region increases in the morning. During 5:00 to 6:00 am, the sill is around 70, which is relatively small, indicating a relatively homogeneous spatial pattern. The variation in population density keeps increasing over time in the morning. During 9:00 to 10:00 in the morning, the sill tripled compared with 5:00 to 6:00 time period, indicating a relatively heterogeneous distribution of population density. Correspondingly, after standardizing the semivariogram, we obtained the correlation in population density across the study region (Figure 4. right side plot). Uniformly, the correlation among locations decrease as distance increase for all time periods. The correlation dropped to zero around 12 km.

**Prediction**

# Discussion: Spatial and Temporal Pattern of Population Density

Based on the Kriging model established above, we displayed the relative population density of each time period during morning peak in Figure 3. Since we aggregated the cellular data in an hourly basis, the estimate of population density could be larger than true value, but it doesn’t affect the relative trend. The Figures show three major spatial processes over time during morning peak in a typical weekday: expanding, growing and converging along traffic corridors (Figure 2). In Shanghai, express ways (dark blue in figure) and metro lines (grey lines in figure) are highly overlapped, making up main transportation corridors that carrying most of daily traffic in Shanghai.

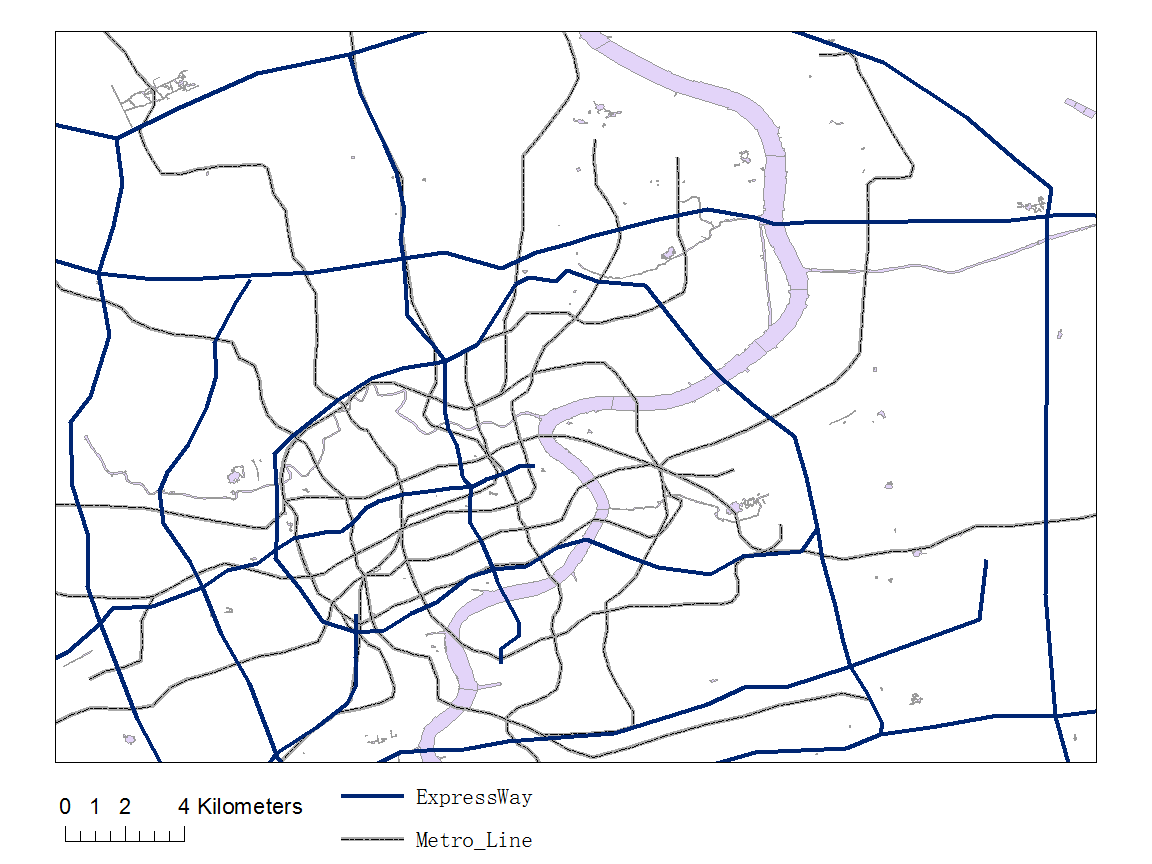


Figure 2. Main Transportation Corridors in Study Region

Early in the morning (5:00 ~ 6:00 am), the population density is relatively homogeneous with density less than 20000 per square kilometer for most area. As shown in the figure, except for the overall uniform pattern, there are multiple weak centers where population density is relatively higher. Among these weak centers, sub centers scattered sparsely around the main center which is much larger than sub centers. The main center located in the middle and extended toward northeast, suggesting relatively high density residential area in the northeast part. In the next hour (6:00 ~7:00 am), the area of centers grew and sub centers began extending towards main center. Population density corridors formed in the process, connecting different centers. This may suggest people began to leave residential area and head to their destinations, creating population density corridors that extend from residential area to other areas. During this process, there is no significant increase in population density within study area.

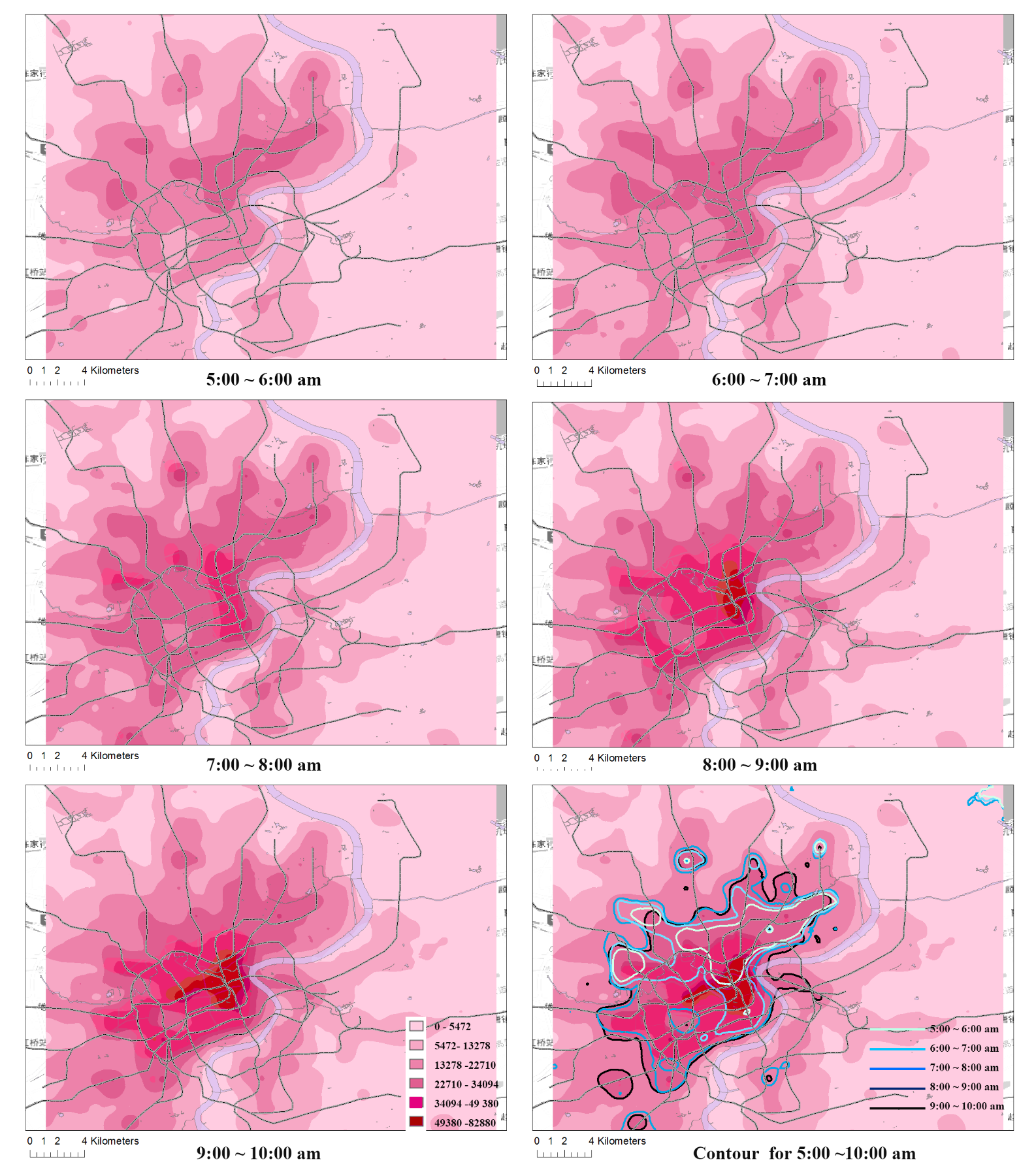


Figure 3. Choropleth Map of Relative Population Density and Contour

From 7:00 to 8:00 in the morning, area of relative high density centers continues to expand while population began to concentrate at some districts, generating new centers of higher population density. The growing of population density continued for the next hour (8:00 ~ 9:00). High density centers in central area keep expanding and connected each other. Significant growth in population density was identified during this period. In central area, the population density doubled to more than 50000 populations per square kilometers. The contour plot shows the expanding trend of relative high population area. Area with population density higher than 22710 pop/km^2 during each hour was outlined in different color. The area expanding process began at 5:00 am, extending from northeast to southwest for three hours and end at 8:00 am. This shows a general southwest moving pattern of population during morning peak. After 8:00 am, the expending process fades while population begins concentrating in the center.

After 9:00 in the morning, sub centers in outer area fade while the population density in central area keeps growing. An observable centripetal tendency of population movement was detected. The population decrease in sub centers suggesting most people has leaving their resident areas at 9:00. While the keep increasing trend of population density in central area reveals the huge attraction force of CBD area. As is shown in the figure, after 9:00 am, there was only one main center and a minor sub center left. At the end of morning peak, population density in central areas has tripled compared to 5:00 in the morning. The main center included the Bund, Nanjing East Road, Nanjing West Road and New World area. The minor sub center is Xujiahui commercial district. Those areas are all major CBD integrating commercial, shopping and entertainment centers. Integrated CBDs are key causes for periodical high traffic pressure. CBDs integrated with shopping center and entertainment center are of great convenience in our daily life. However, the tremendous traffic attraction exerts great periodical pressure to the local transportation system.