# Spatio-temporal models for floating population in Seoul

2020321804 Seung ji Nam

December 6, 2021

## 1 Introduction

The objective of this project is to analyze the floating population in Seoul. There are IoT sensors called Smart Seoul city data sensor, S-Dot in Seoul and each of the sensors has a different usage such as monitoring PM(particulate matter), urban wind path, and other environmental indicators. There are 93 sensors that measures floating populations every 10 minutes in Seoul since May 2020 and the data is available from the open data platform website of the Seoul Metropolitan Government https://data.seoul.go.kr/dataList/OA-15964/S/1/datasetView.do. According to the website, the sensors are installed throughout Seoul, but especially near traditional markets or tourist spots.

Here, data from January 2021 to October 2021 is used. Also, there are five other variables to predict and understand the floating population. Other variables explain population, building, and transportation usage, and they are also obtained from the same website above.

### 1.1 Data

Data uses floating population observed at 93 different sensors placed in Seoul and the data of every 10 minutes is summarized into a sum of floating population per day. Considering missing value of data, data from 4th January to 23rd October is analyzed. Other variables are calculated based on the meaning of the variable. Lee et al.(2019) proposed that the influence of bus stops and subway stations is 300m and 500m. Also, it is known that the neighborhood unit or living area range is 500m distance.

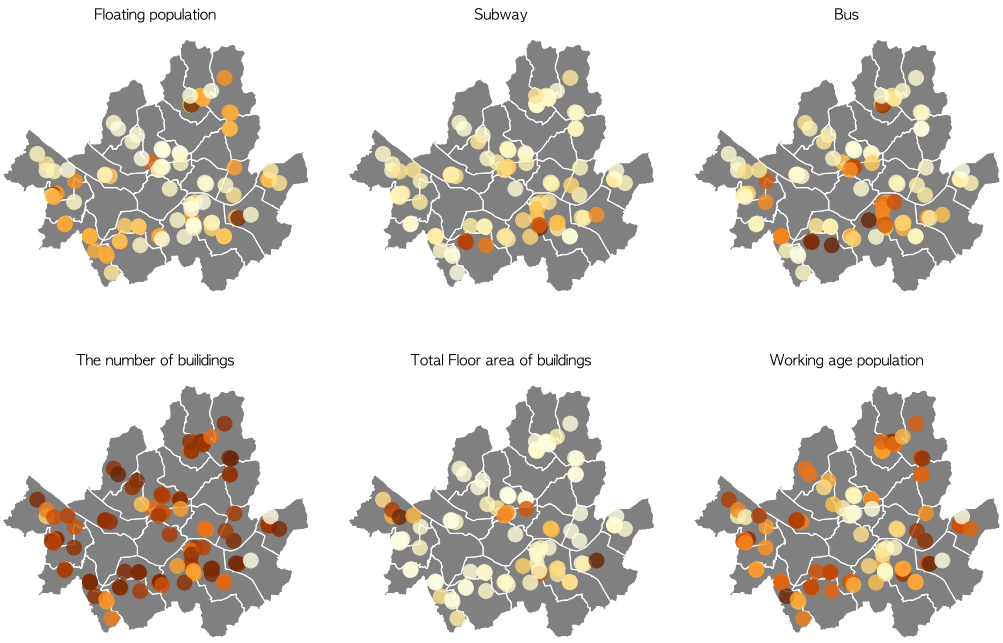


Figure 1: The color of points indicates the value of each variable. The plot is based on average data in October.

In summary, the number of people getting on and off the bus stop within 300m of sensors is calculated. In the same way, the number of people using subway stations within 500m of sensors is considered. The number of buildings, the total floor area of buildings, and working-age population within the 500m distance from the sensor are counted. Subway and bus variables are daily data and other variables have the same value in the total period.

Since the floating population is collected at different sensor points, the type of data is point-referenced data. It can be easily supposed that the data has both spatial and temporal dependence structures. Therefore, we check if there are spatial and temporal correlations among the values of the floating population.

# 2 Spatio-temporal dependence

## 2.1 Spatial correlation

We used only March and October data evaluating spatial data to remove expected temporal correlation. Without considering spatial dependency, we can estimate $\beta$ coefficients using linear regression model. Then, we can suppose there are spatial correlation left in the residual of the model and we can fit variogram with the residuals, which enables to fit the covariance function. We observe as distance gets larger, estimated variogram shows different pattern and this also can be checked from Figure 2.

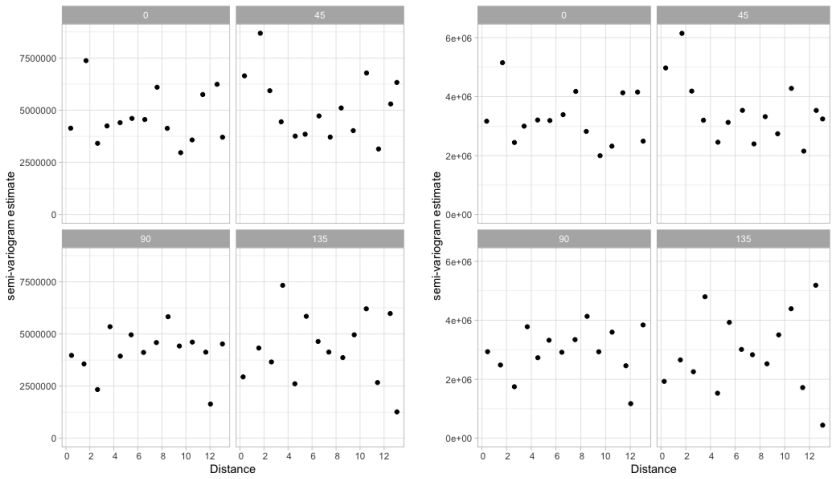


Figure 2: *Left*: Variogram of floating population in March, *Right*: Variogram of floating population in October

But variograms has limitation.

## 2.2 Temporal correlation

We checked if there is a temporal dependency in the data on a daily, weekly, and monthly basis. Here, we use two methods; variogram for spatio-temporal data and Auto-correlation plot. For spatio-temporal data, it is possible to use similar variogram estimation in 2.1. Assuming separable and nonseparble covariance between time and space, we can detect temporal, spatial correlation, and correlation between time and space.

From the variogram in Figure 3, we could say that time and space covariances are separable because it is hard to see pattern between space lag and time lag. Because building and population variables have the same value for all period, when fitting variogram and covariance function, we only use subway and bus variables. Also, temporal correlation cannot be strongly detected from Auto-correlation plot. Two methods are done in all daily, weekly, and monthly setting. Though daily data shows a slight temporal correlation compared to weekly and monthly data, it does not imply temporal dependency strongly. I concluded that temporal correlation does not have to be considered especially on weekly and monthly

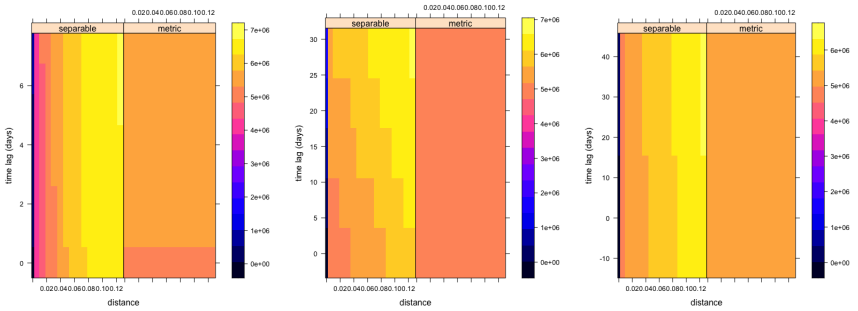


Figure 3: From left, variogram for spatio-temporal data of daily, weekly, and monthly data

based data. Still, there might be a possibility that daily based data has a temporal dependency given that sensors are located around tourist places and markets and floating population depends on day of the week.

# 3 Methodology

Hierarchical spatial model SGLMM is a model that extends GLM to include spatial random process. There are two advantages of SGLMM; it is simple to set spatial process as a random effect conditionally and the expectation of data can be modeled using link function.

Suppose $\mathbf{Y} = (Y_1, ..., Y_n)$ are floating population at each censor and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ indicate independent variables which are the number of people using subway and bus, the number of buildings, the total floor area of buildings and working age population. Spatial process of 93 sensors is assumed as a Gaussian random effect Z(s). Then, $\mathbf{Y}$ follows poisson distribution given $\boldsymbol{\beta}$ and spatial random effect $Z$ and models can be defined as below.

$$\begin{aligned} \mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta} &\sim Poisson(exp(\boldsymbol{X\beta} + \boldsymbol{Z})) \\ Z(s)|\sigma^2, \rho &\sim N(0, \sigma^2\Gamma(\rho)) \\ \boldsymbol{\beta} &\sim N(0, 10\mathbf{I}) \\ \sigma^2 &\sim InverseGamma(0.1, 0.1) \\ \rho &\sim U(0, 1) \end{aligned} \tag{1}$$

Since the model does not include temporal factors, model is conducted to average data per month and I compare the coefficients from model of each month. In addition, Bayesian inference MCMC method is applied to estimate $\boldsymbol{\beta}$ coefficients, $\sigma$, and $\rho$. Therefore, convergence of each estimation and MCMC results are evaluated. For calculating MCMC algorithm, NIMBLE function in R is used.

# 4 Result

The same SGLMM model is applied to March, May, July, and September 2021. Most of MCMC result converges well as below. (Figure 4, Table 1  2 ). The number of iterations is 200,000 for each model, the number of chains is used 5 time, and the result is chosen based on convergence. The number of burn-in is 5000. All models have acceptance rate at least 0.439.

The lower and upper bound of HPD 95% Interval of $\beta_1$, $\beta_2$, and $\beta_5$ are not included in the same sign and these three variables show slightly different values in each month compared to the result of other months. The number of building, and the total floor area of building withing 500m from sensors shows positive correlation with floating population in overall months.

| | Posterior mean | HPD 95% Interval | Acceptance Rate |
|---|---|---|---|
| $\beta_1$ (subway) | 0.458 | (-0.439, 1.409) | 0.440 |
| $\beta_2$ (bus) | 0.058 | (-1.224, 1.218) | 0.439 |
| $\beta_3$ (# of building) | 1.769 | (1.102, 2.449) | 0.440 |
| $\beta_4$ (total floor area) | 0.813 | (0.024, 1.581) | 0.440 |
| $\beta_5$ (population) | -0.235 | (-1.105, 0.615) | 0.440 |
| $\rho$ | 0.006 | (0.000, 0.017) | 0.448 |
| $\sigma^2$ | 0.633 | (0.333, 0.991) | 0.445 |

Table 1: MCMC result of October data

| | March | May | July | October |
|---|---|---|---|---|
| $\beta_1$ (subway) | 0.592 | 0.274 | 0.267 | 0.458 |
| $\beta_2$ (bus) | -0.320 | -0.217 | -0.277 | 0.058 |
| $\beta_3$ (# of building) | 1.872 | 1.806 | 1.886 | 1.769 |
| $\beta_4$ (total floor area) | 1.002 | 0.770 | 0.763 | 0.813 |
| $\beta_5$ (population) | -0.285 | 0.100 | 0.001 | -0.235 |
| $\rho$ | 0.003 | 0.004 | 0.002 | 0.006 |
| $\sigma^2$ | 0.546 | 0.459 | 0.464 | 0.633 |

Table 2: Estimated posterior mean from MCMC results in March, May, July, and October

# 5 Conclusion

At first, I assumed there will be a strong temporal correlation between floating population. However, the correlation analysis did not show strong dependency. Therefore, SGLMM model is implied to understand floating population in city. Though all the results throughout the month are not same, the estimate of each variable in each month is converged well so that we could interpret the result. I suppose there is a limitation in data. If floating population data is collected evenly throughout Seoul, I suppose the data will have more spatial and temporal correlation and other variables will be more significant. For further research, I suggest to do temporal spatial glmm model using daily data or every 10 minutes data. Though correlation between time and space is separable shown in 2.2, covariance matrix gets really bigger when assuming time and space effect as a random effect and computation might be heavy.

# References


1. Man Ho Lee, Jong Hoon Lee, Ho Sun Woon, and Eui Young Shon (2019). "A Study on the Improvement of Bus Traffic Assignment Considering Catchment Area and Access Distance by Bus-Stop" Seoul Studies, 20(3) 79-90.

2. Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics, Journal of the Royal Statistical Society: Series C (Applied Statistics), 47, 299-350.

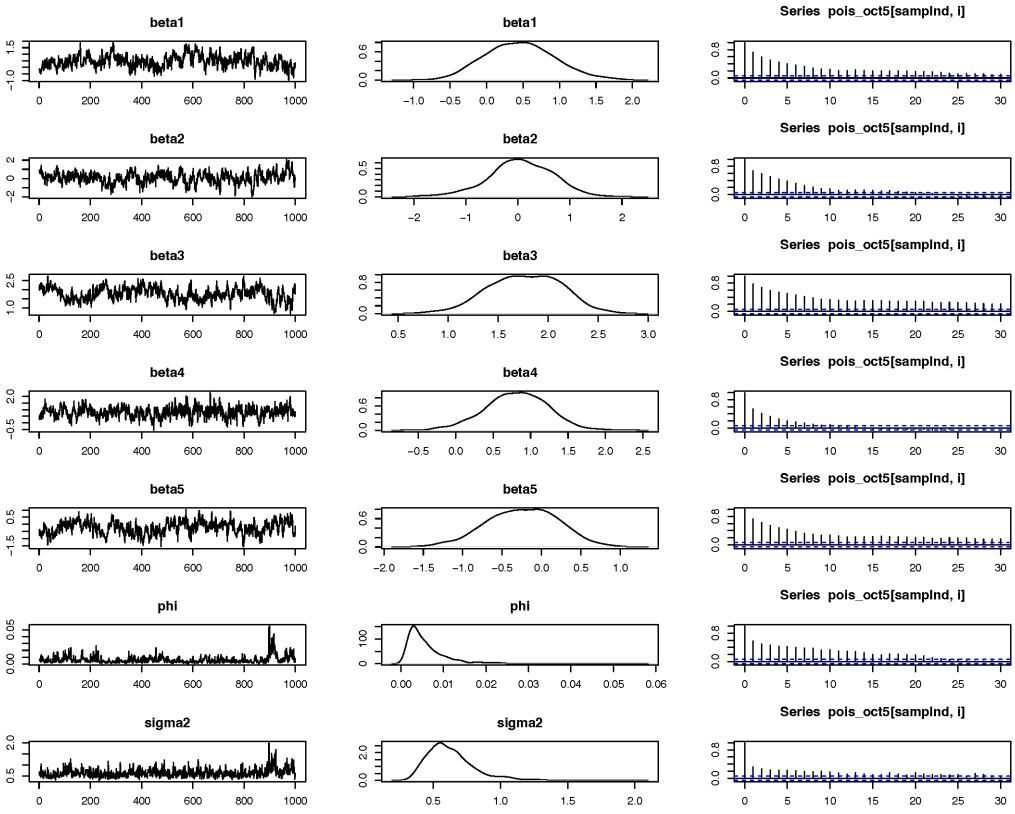


Figure 4: MCMC result of October data