Spatial pattern of population movement during morning peak hours

STAT 5544 Final Report

Wenyu Gao, Danni Lu

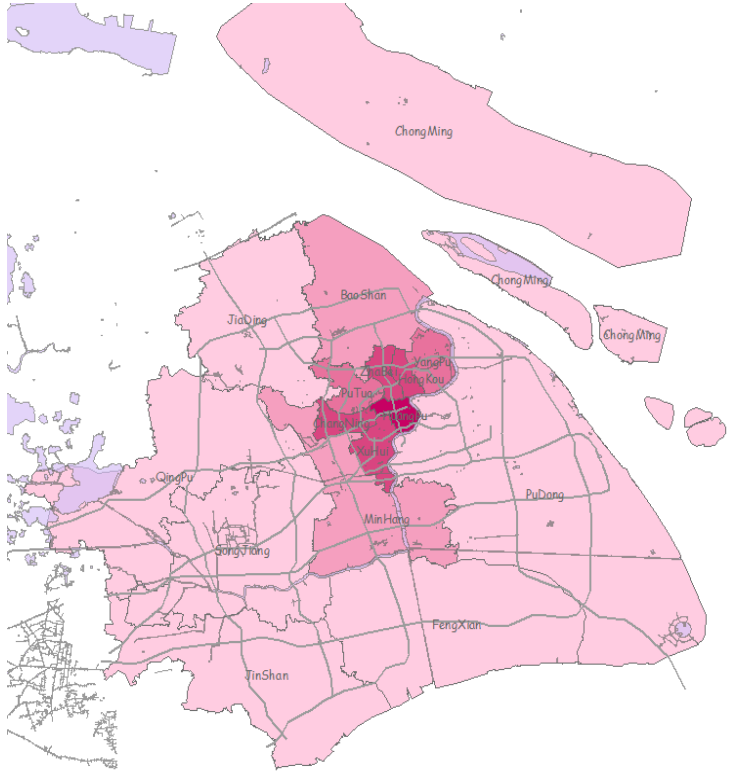
Dec.1st, 2016

CONTENTS

CONTENTS

INTRODUCTION

- Data source



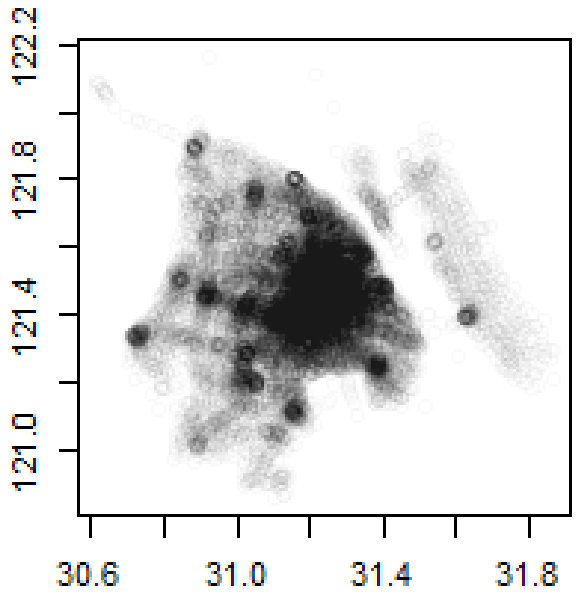
Shanghai

Area
- Municipality 6,341 km^2

Population (2015)
- Municipality 24 million
- Rank 1st in China
- Density 3,800/km^2

Cell Tower Data

Cell towers: 37,450
Density: 5.91 /km^2
Time: consecutive 24 hour
Records: 1.1 billion (80 GB)



INTRODUCTION

- Original Dataset

ID	LONG	LAT	DATE	TIME	EVENT
FEXIN123998659 82DF	121.3535	31.13478	2014/4/3	0:56:21	12
ASKXIN1S23SDSD9986	121.3644	31.13478	2014/4/3	12:25:07	10
ASDSD2312KVHOJSD12	121.3753	31.13478	2014/4/3	11:34:11	10

- Privacy
  - ✓ Records are anonymized
  - ✓ Records are aggregated by areal unit and time period
  - ✓ No personal information is involved and displayed

INTRODUCTION

- Research question

Where do people go during morning rush hours?

How does the pattern change over time ?

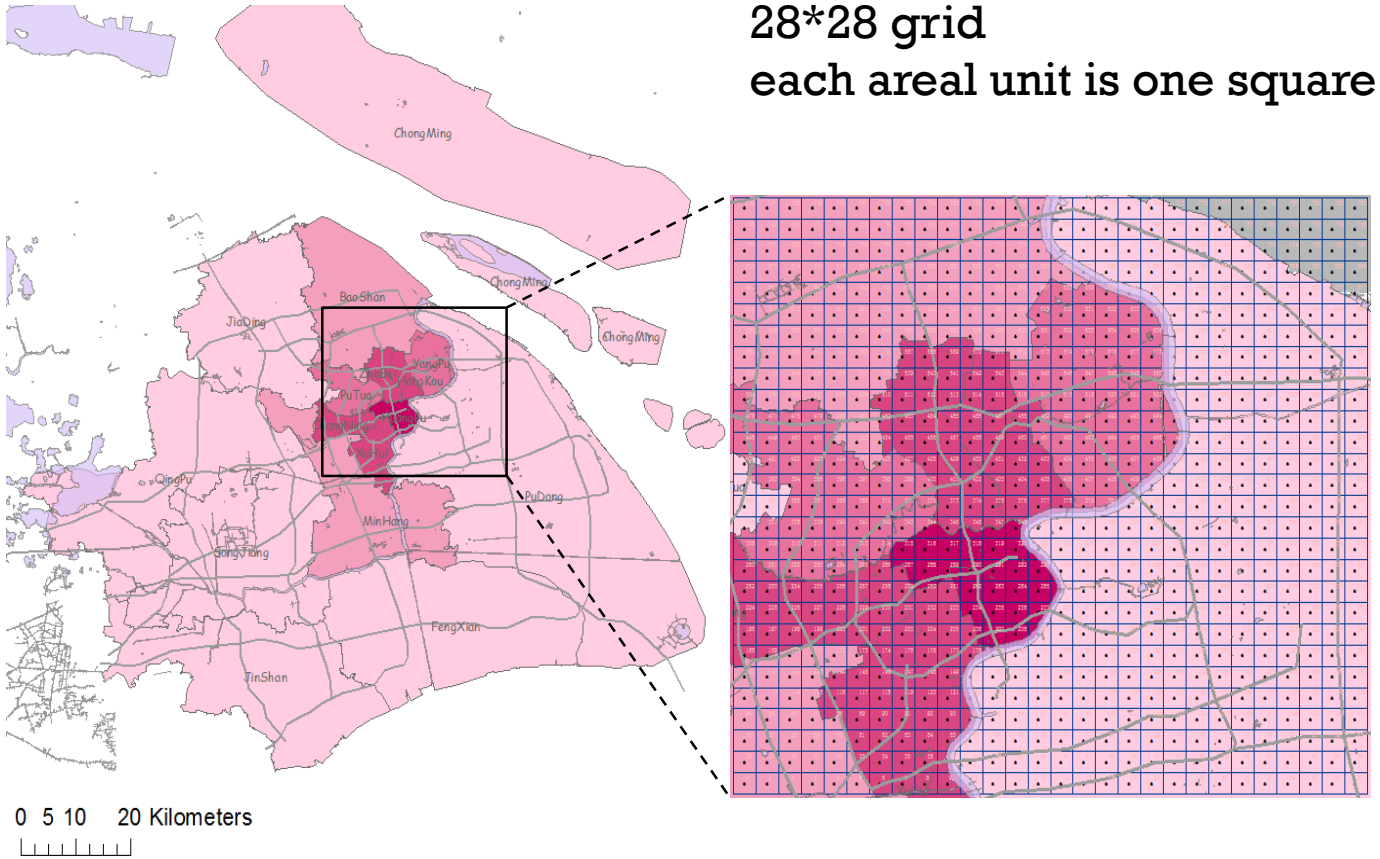
CONTENTS

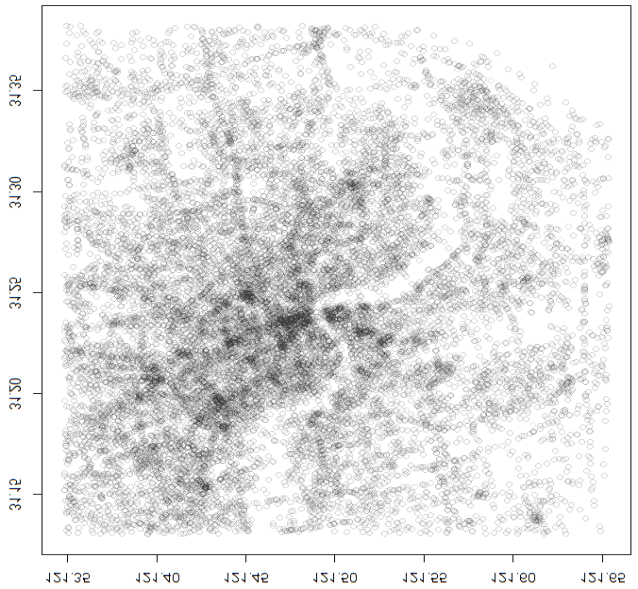
EXPLORATORY DATA ANALYSIS

- Data aggregation

Urban area

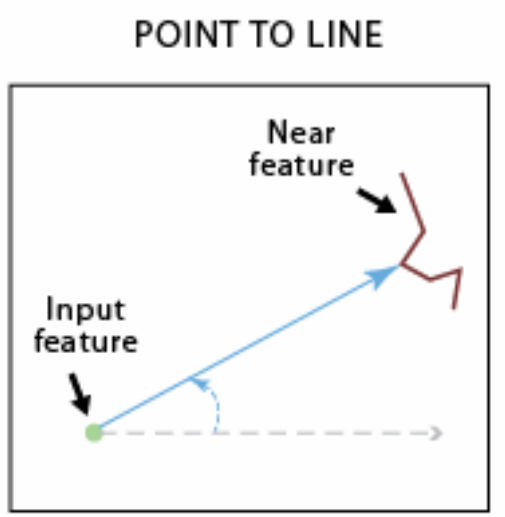
28*28 grid
each areal unit is one square kilometer





EXPLORATORY DATA ANALYSIS

- Data aggregation
- Calculate covariates:
  - Distance to metro line
  - Distance to expressway



Longitude

```
Min.   :121.4
1st Qu.:121.4
Median :121.5
Mean   :121.5
3rd Qu.:121.6
Max.   :121.6
```

Latitude

```
Min.   :31.13
1st Qu.:31.20
Median :31.26
Mean   :31.26
3rd Qu.:31.32
Max.   :31.38
```

Density

```
Min.   :    0
1st Qu.: 2060
Median : 5894
Mean   : 9773
3rd Qu.:14575
Max.   :83822
```

D2Metro

```
Min.   :   0.395
1st Qu.: 387.974
Median : 973.144
Mean   :1373.922
3rd Qu.:1981.959
Max.   :6852.305
```

D2Road

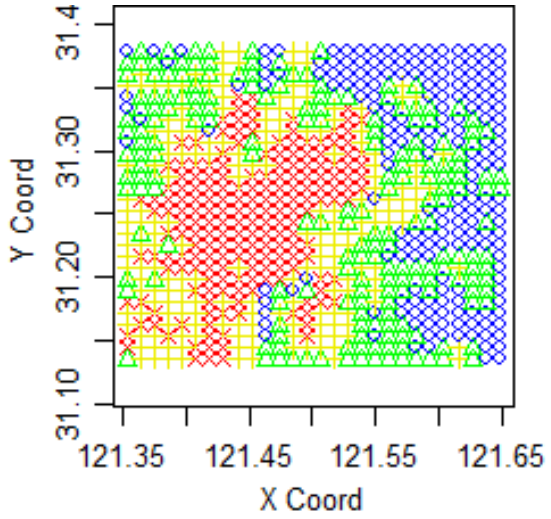
```
Min.   :   2.274
1st Qu.: 397.764
Median : 888.332
Mean   :1126.451
3rd Qu.:1609.672
Max.   :5607.713
```

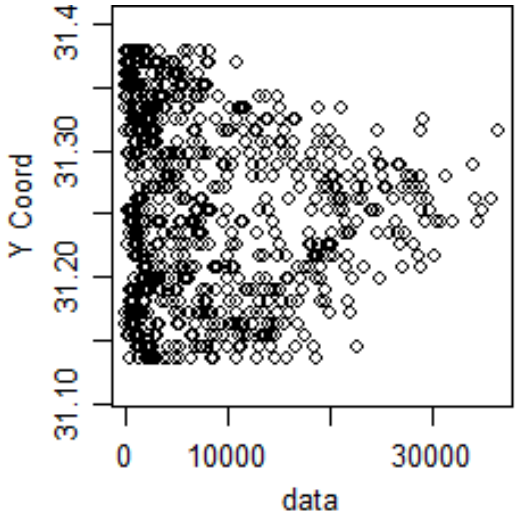
Hour

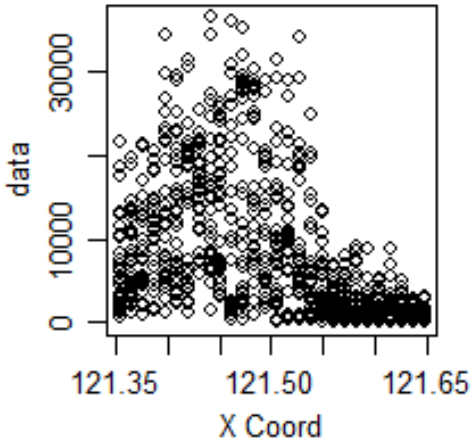
```
Min.   :5
1st Qu.:6
Median :7
Mean   :7
3rd Qu.:8
Max.   :9
```

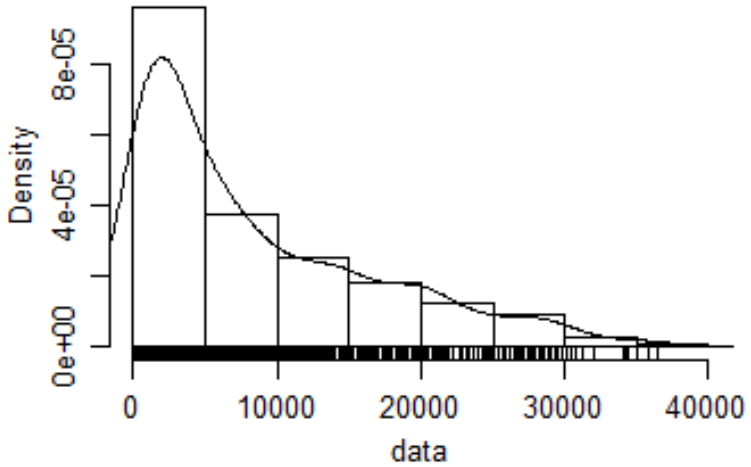
EXPLORATORY DATA ANALYSIS

Population density at 5:00-6:00 am



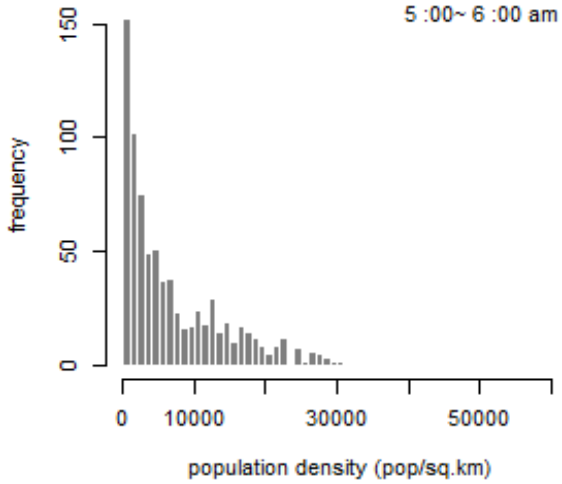


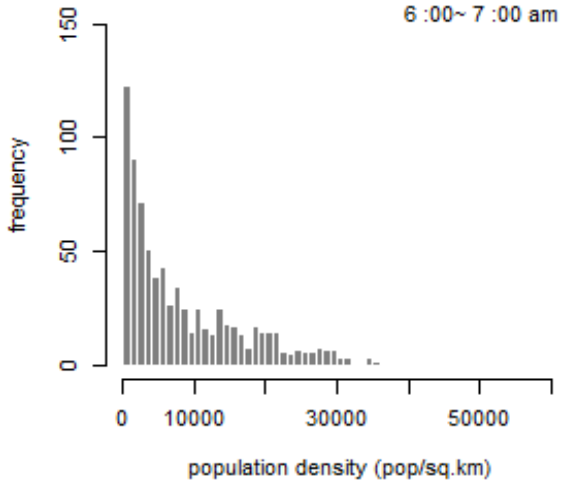


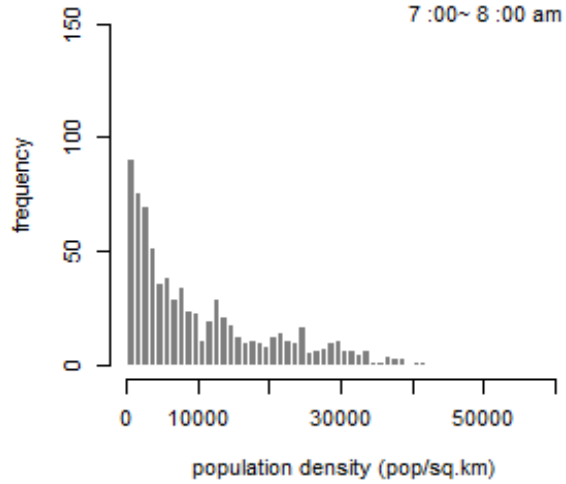


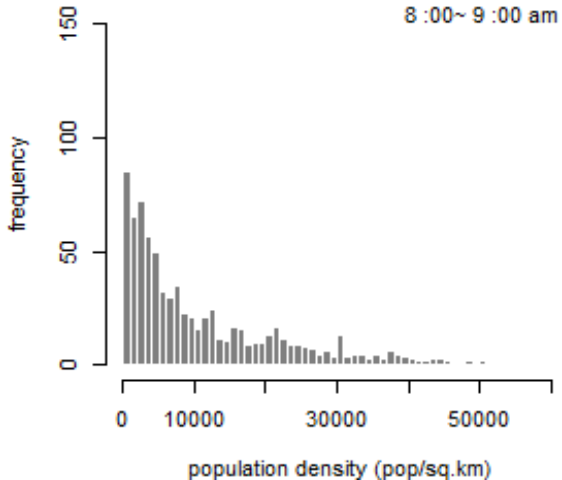
EXPLORATORY DATA ANALYSIS

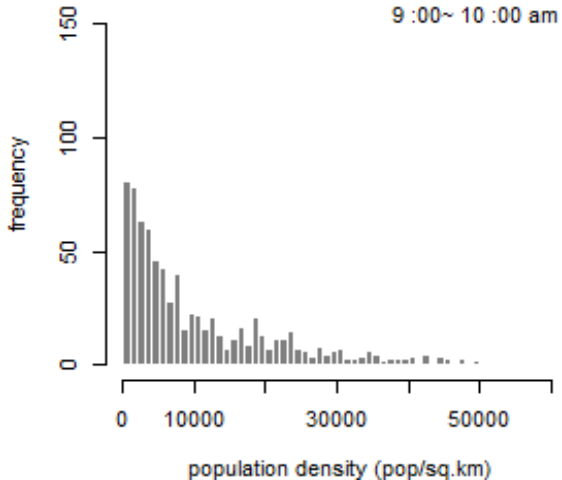
Population density at each hour







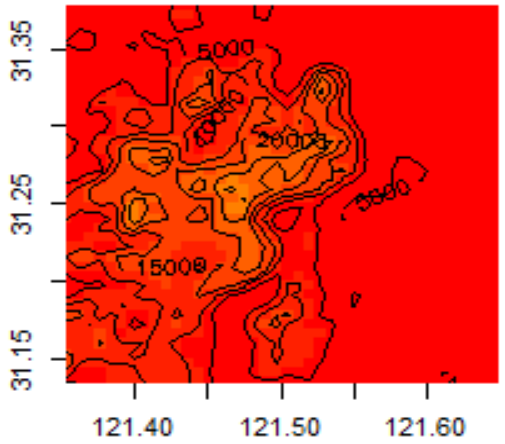




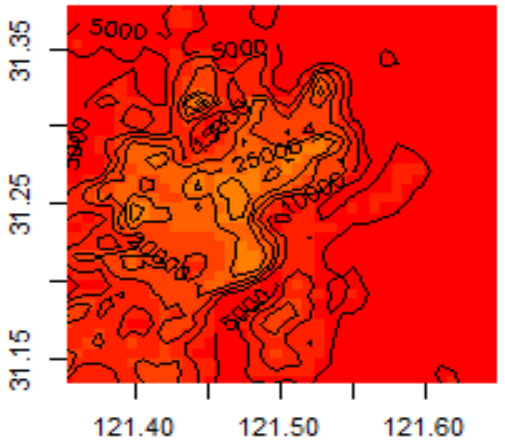
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5:00-6:00	0	1378	4217	7003	11200	33550
6:00-7:00	0	1849	5322	8421	13430	36390
7:00-8:00	0	2452	6940	10550	16050	47720
8:00-9:00	0	2588	7033	11440	16990	68800
9:00-10:00	0	2554	6772	11450	17050	83820

EXPLORATORY DATA ANALYSIS

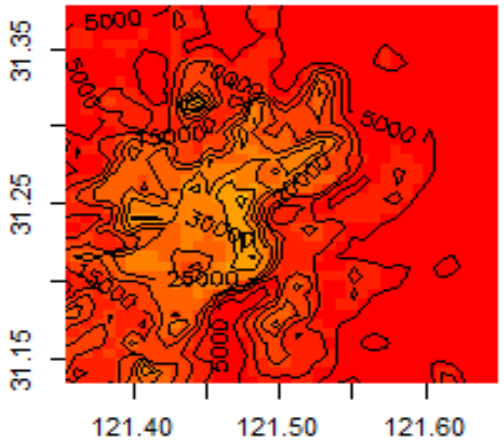
Population density at each hour



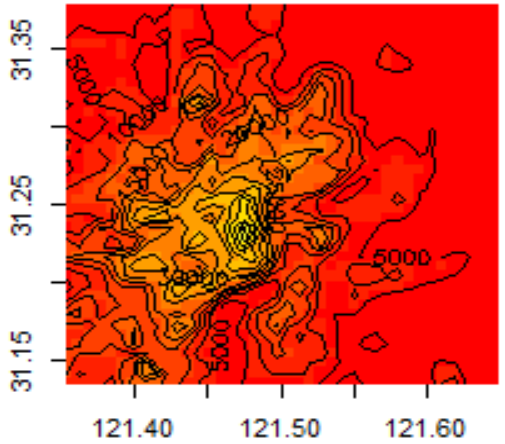
5 : 00 ~ 6 : 00 am



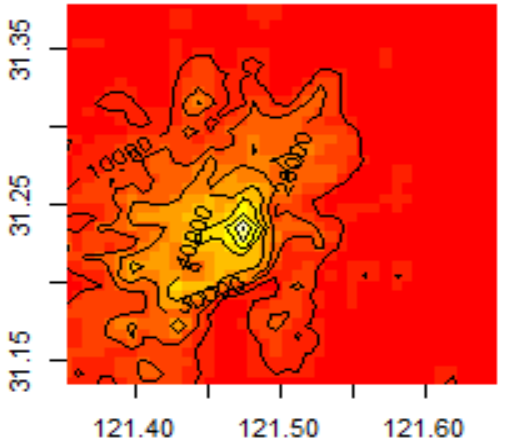
6 : 00 ~ 7 : 00 am



7 : 00 ~ 8 : 00 am



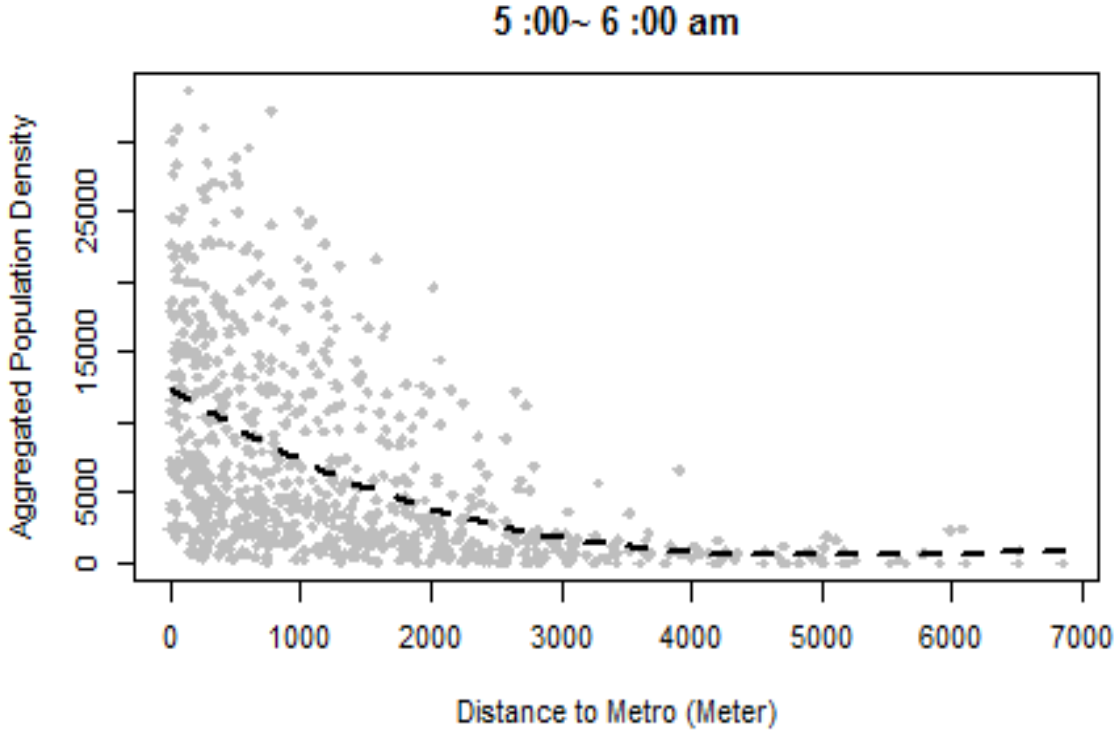
8 : 00 ~ 9 : 00 am

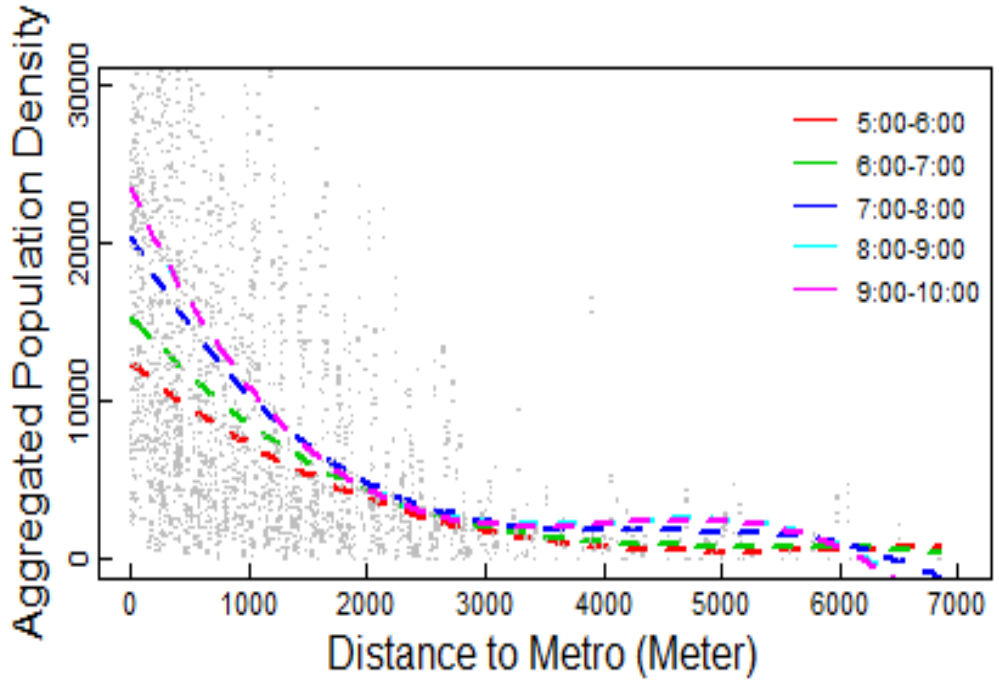


9 : 00 ~ 10 : 00 am

EXPLORATORY DATA ANALYSIS

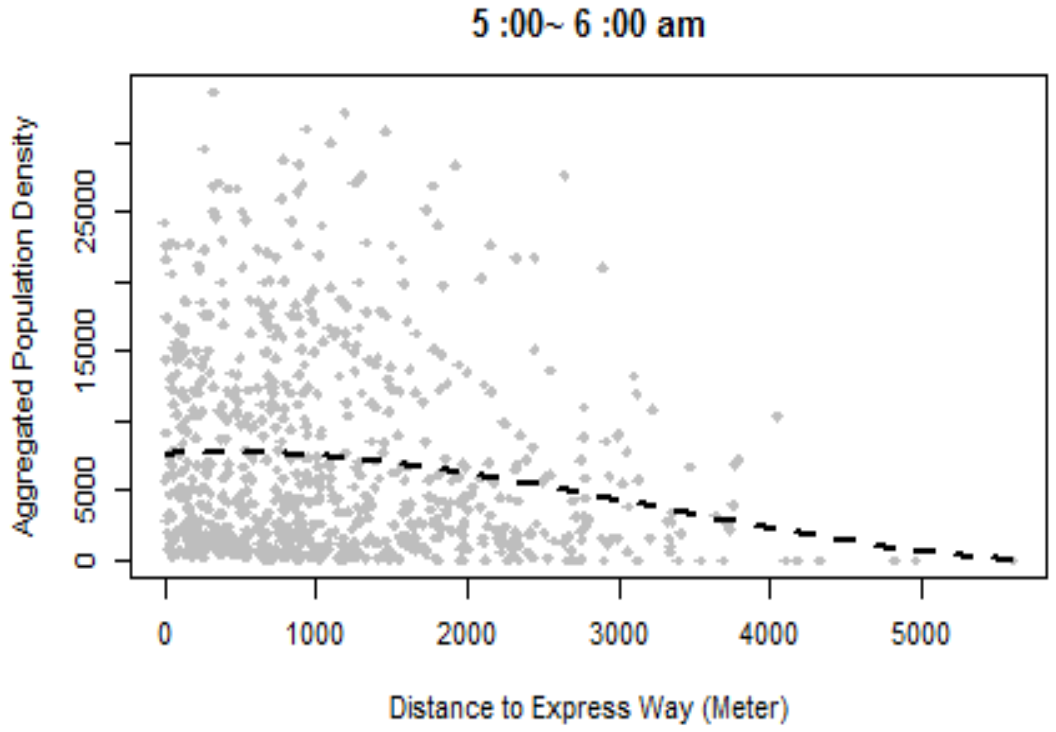
- Covariates – Distance to Metro Line

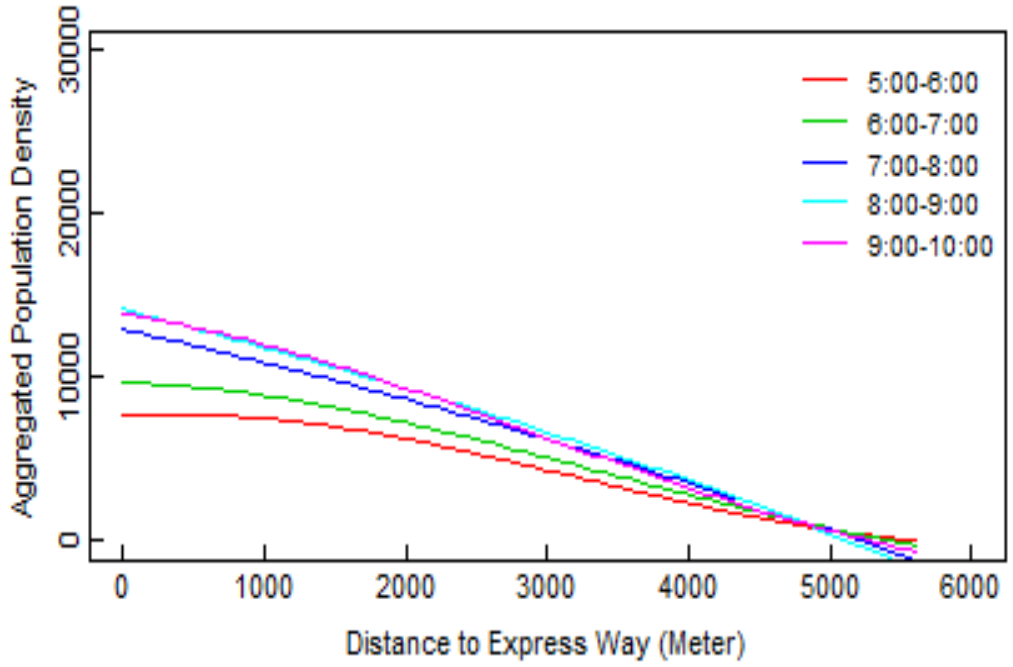




EXPLORATORY DATA ANALYSIS

- Covariates – Distance to Expressway

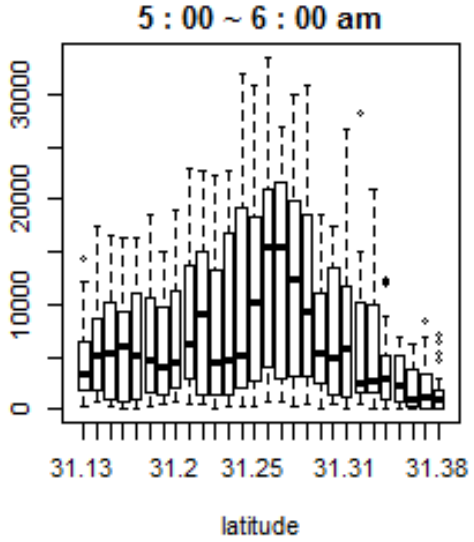


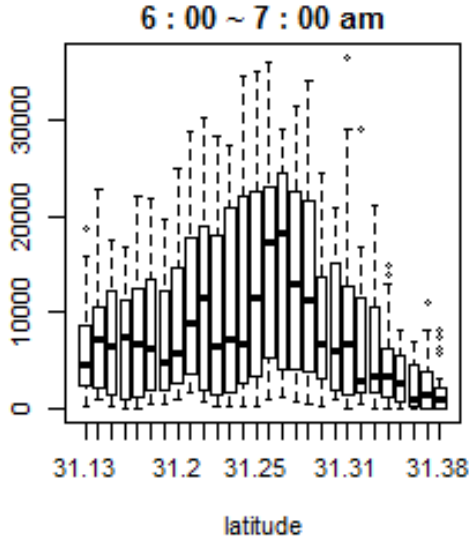


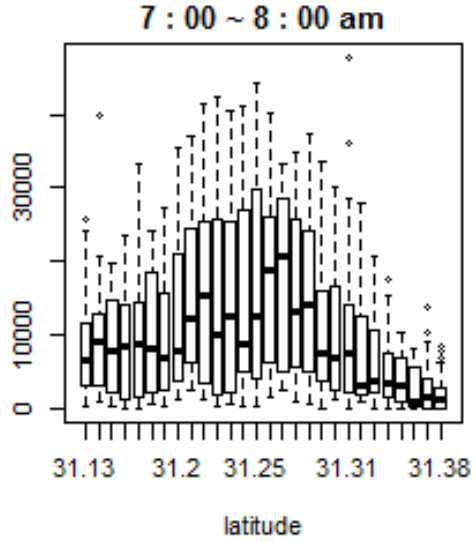
CONTENTS

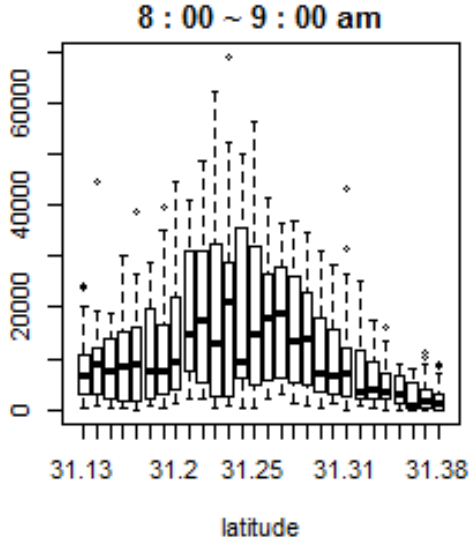
SPATIAL TREND

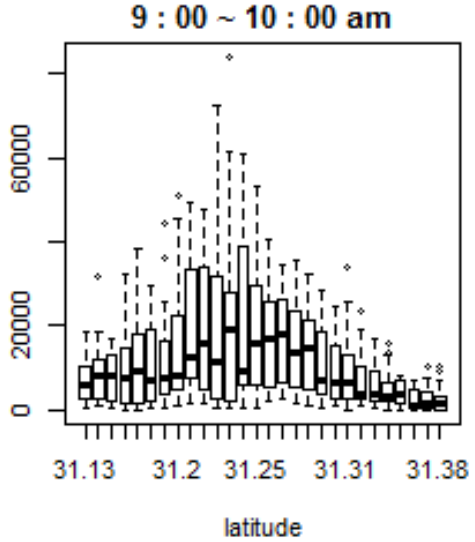
- Latitude





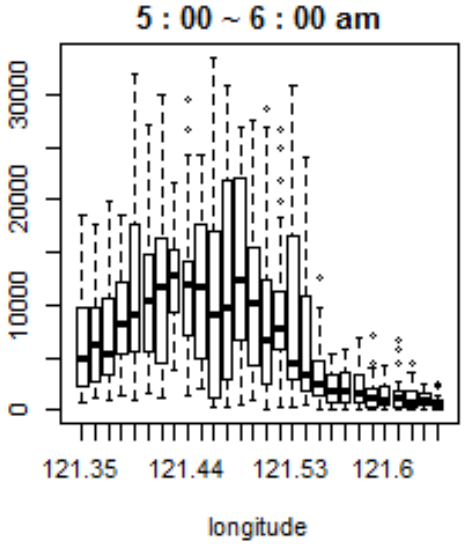


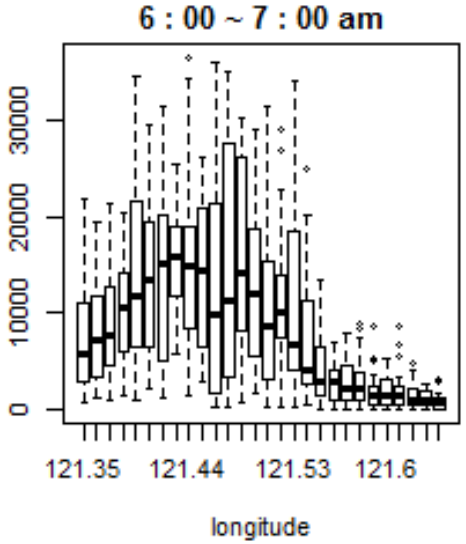


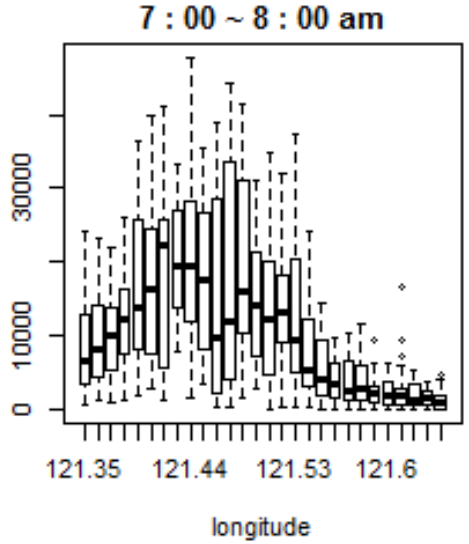


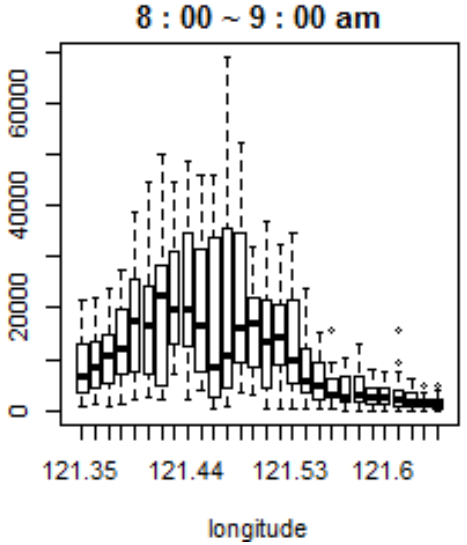
SPATIAL TREND

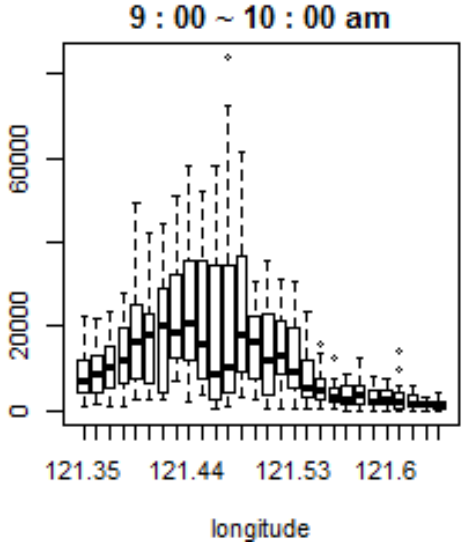
- Longitude











SPATIAL TREND

- Detrend

  $Y(s_i)$: *Population density at aggregated cell tower* s_i;

  x_{1i}: *latitude of site* (s_i);

  x_{2i}: *longitude of site* (s_i);

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

- Stepwise selection (AIC)

  ✓ Full model selected

  ✓ Define new response as residual from full model

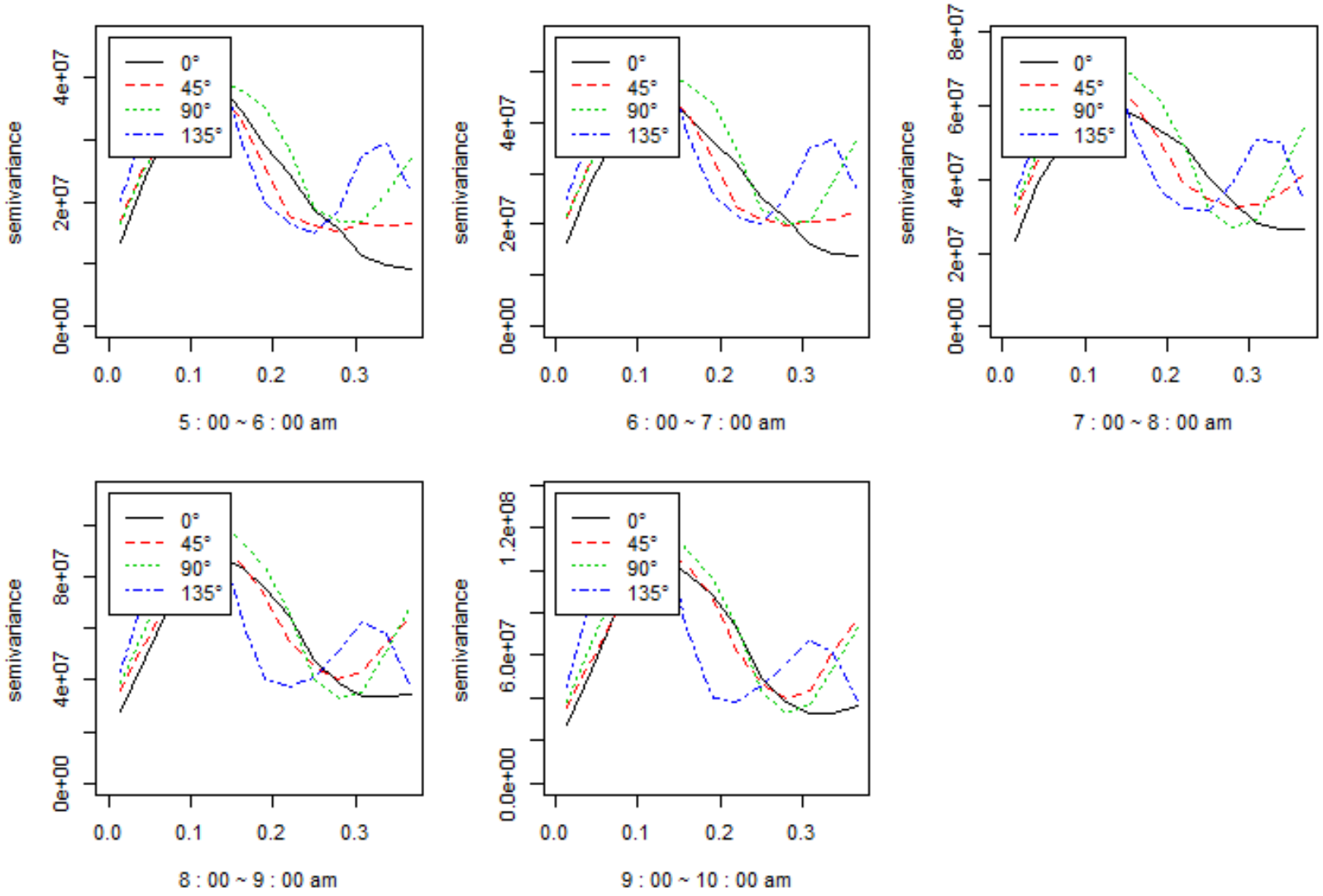
  $$Y(s_i) = Y(s_i)_{original} - \widehat{Y(s_i)}$$

```
> summary(fit1)

Call:
lm(formula = data ~ .^2 + I(long^2) + I(lat^2), data = gdata5)

Residuals:
   Min     1Q Median     3Q    Max
-11460  -3852   -555   2726  20568

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.524e+09  4.241e+08 -10.669  < 2e-16 ***
long         6.932e+07  6.759e+06  10.256  < 2e-16 ***
lat          2.018e+07  4.428e+06   4.556 6.05e-06 ***
I(long^2)   -3.018e+05  2.755e+04 -10.954  < 2e-16 ***
I(lat^2)    -5.698e+05  4.062e+04 -14.030  < 2e-16 ***
long:lat     1.271e+05  2.986e+04   4.255 2.35e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5361 on 778 degrees of freedom
Multiple R-squared:  0.4494,    Adjusted R-squared:  0.4458
F-statistic:   127 on 5 and 778 DF,  p-value: < 2.2e-16
```

CONTENTS

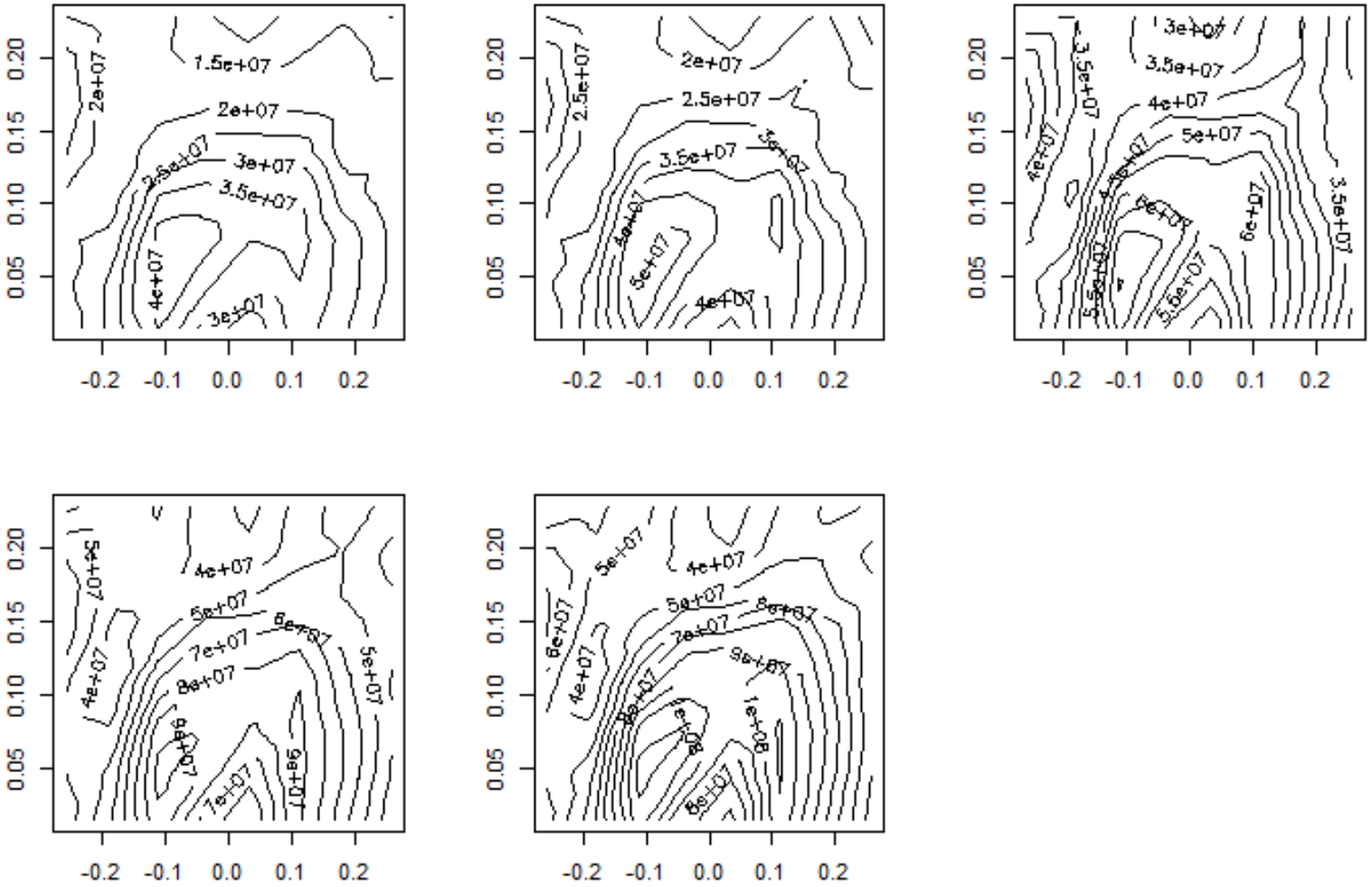
ANISOTROPY

- Directional semivariogram



ANISOTROPY

- Contour Plot

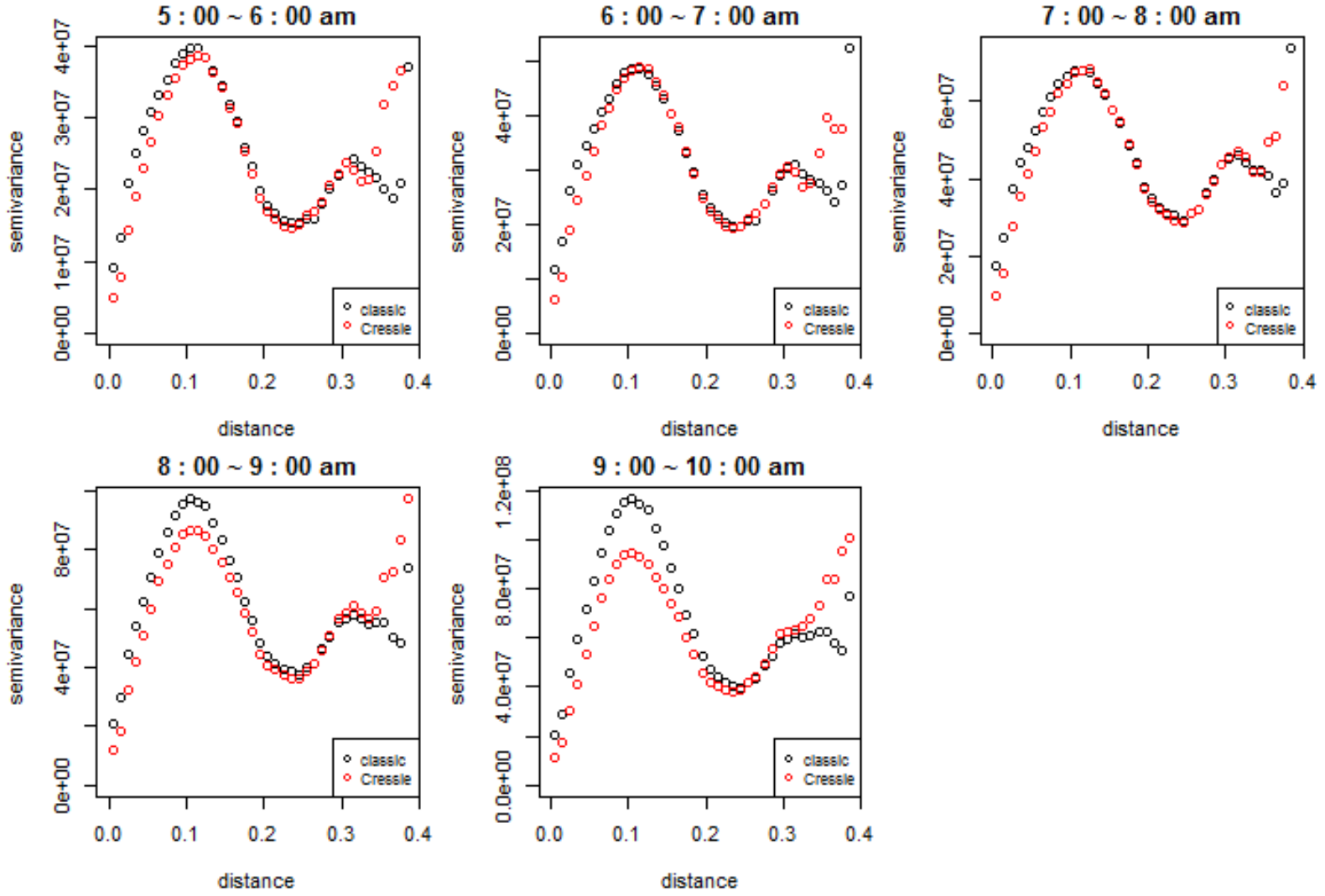


CONTENTS

MODEL FITTING

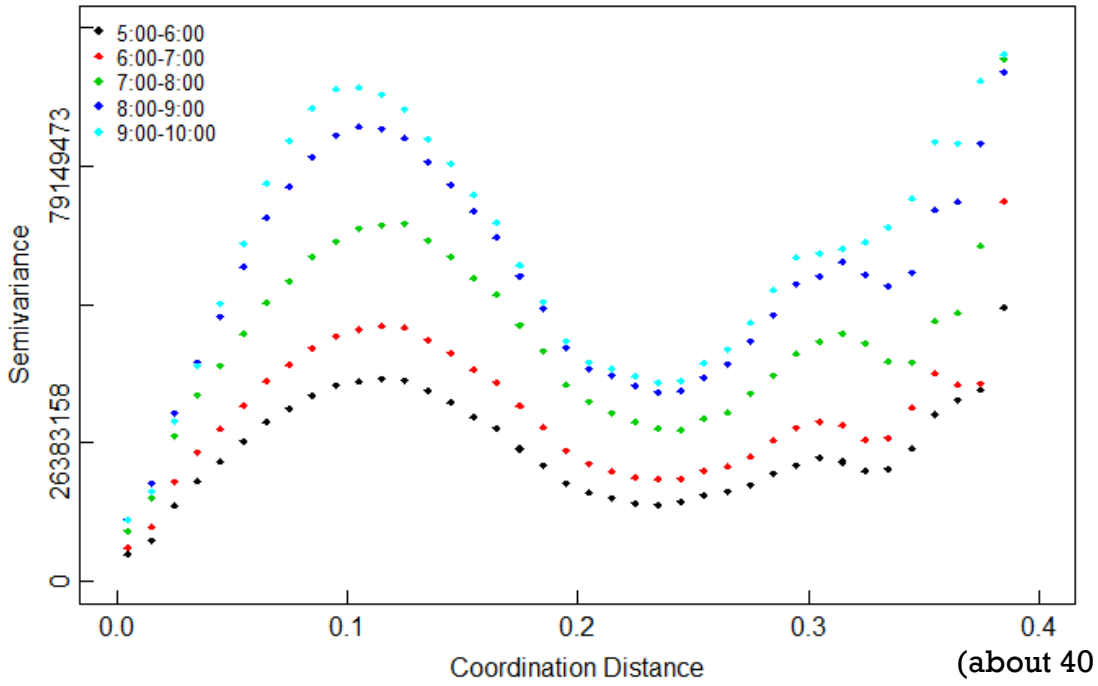
- Semivariogram

Classical & Cressie

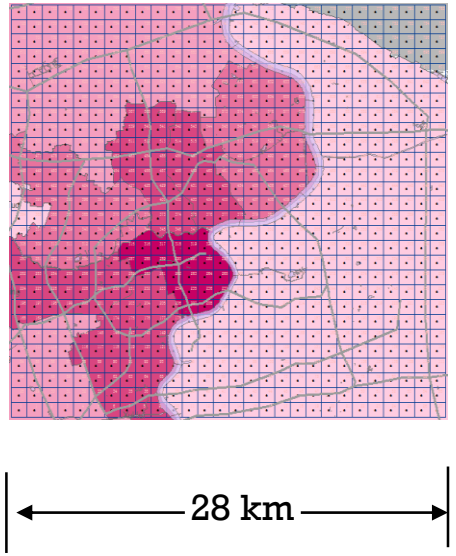


MODEL FITTING

- Semivariogram



(about 40 km)



Trend: Homogenous to heterogeneous distribution of population density

MODEL FITTING

- Semivariogram fitting – WLS with Cressie weight

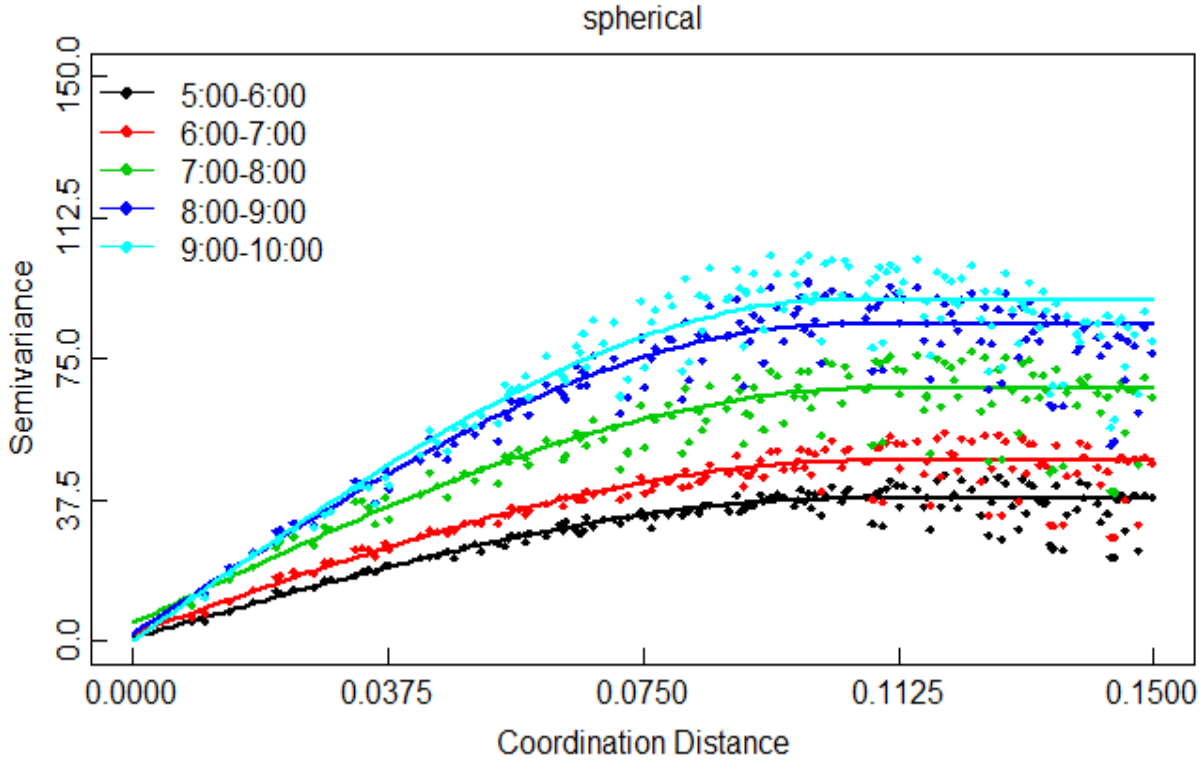
Residuals for fitted model:

	5	6	7	8	9	mean
Spherical	1345.829	1492.512	1584.260	1087.853	1313.328	1364.756
Exponential	1784.886	1777.499	1737.484	1778.246	2688.834	1953.390
Gaussian	1585.495	1768.071	1850.299	1292.079	1369.624	1573.114
Cubic	1564.827	1744.105	1826.269	1236.830	1278.881	1530.182
Matern	1784.886	1777.499	1737.484	1778.246	2688.834	1953.390
Circular	1388.799	1544.079	1635.957	1082.567	1200.924	1370.465
Power	4369.290	4148.123	3914.707	4643.682	6112.160	4637.592
Powered.exponential	3177.101	3017.073	2784.187	3302.948	4695.429	3395.347

MODEL FITTING

- Semivariogram fitting – WLS with Cressie weight

Spherical model:




```
> semfit_s
  nugget sigma.sq     phi   range sum.of.sq
5 0.8011  36.9817 0.1054 0.1054  1345.829
6 1.9464  45.9253 0.1074 0.1074  1492.512
7 4.7192  62.3882 0.1086 0.1086  1584.260
8 1.9597  82.1135 0.1037 0.1037  1087.853
9 0.0000  90.8891 0.1047 0.1047  1313.328
```

MODEL FITTING

- Likelihood method

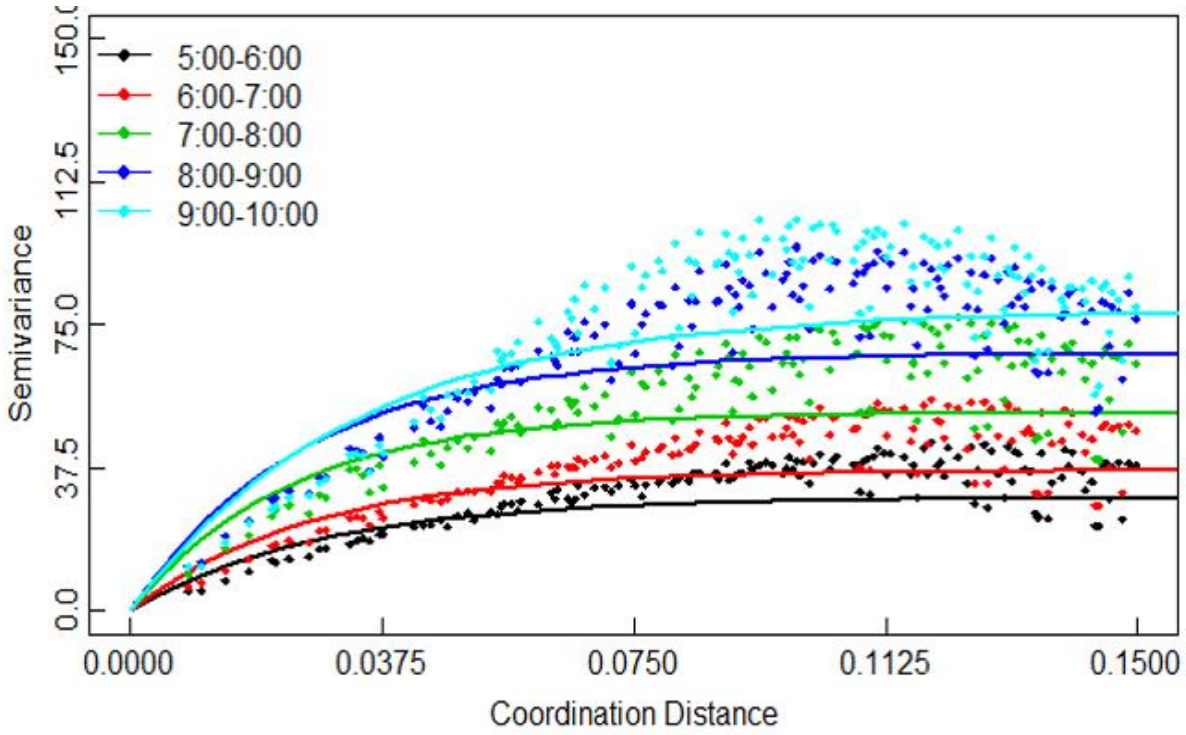
AIC and BIC for fitted model:

	5	6	7	8	9	mean
Spherical AIC	4132.694	4342.548	4689.473	4830.251	4771.841	4553.361
Exponential AIC	4108.039	4315.906	4659.611	4805.030	4756.671	4529.051
Gaussian AIC	4859.872	5033.190	5318.416	5550.959	5661.542	5284.796
Cubic AIC	4123.605	4336.645	4684.590	4854.445	4831.221	4566.101
Matern AIC	4108.039	4315.906	4659.611	4805.030	4756.671	4529.051
Spherical BIC	4151.352	4361.206	4708.131	4848.909	4790.498	4572.019
Exponential BIC	4126.696	4334.564	4678.268	4823.687	4775.328	4547.709
Gaussian BIC	4878.530	5051.848	5337.074	5569.617	5680.200	5303.454
Cubic BIC	4142.262	4355.302	4703.247	4873.103	4849.879	4584.759
Matern BIC	4126.696	4334.564	4678.268	4823.687	4775.328	4547.709

MODEL FITTING

- Likelihood method

Exponential and Matern model:



	nugget	sigma.sq	phi	range	AIC	BIC
5	0	29.7172	0.0280	0.0838	4108.039	4126.696
6	0	36.8529	0.0264	0.0791	4315.906	4334.564
7	0	52.0373	0.0236	0.0707	4659.611	4678.268
8	0	67.6768	0.0259	0.0776	4805.030	4823.687
9	0	78.7829	0.0331	0.0991	4756.671	4775.328

MODEL FITTING

- Relation with covariates -- GAM model with Gamma family

$$Y(s_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_5 x_{1i} x_{2i} + S(lat, long) + \epsilon_i;$$

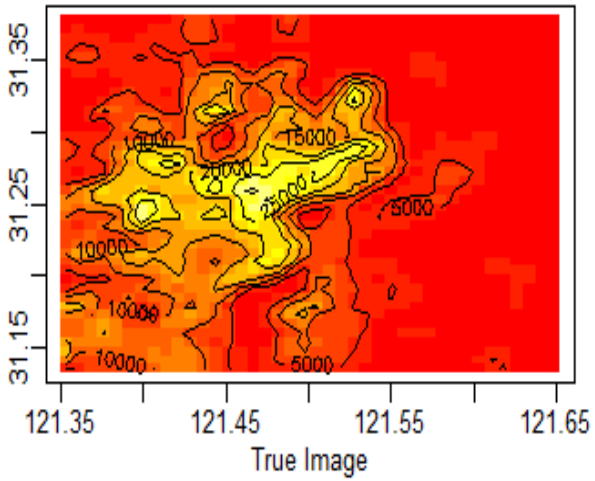
$Y(s_i)$: *Population density at aggregated cell tower* s_i;
x_{1i}: *distance to Metro at site* (s_i);
x_{2i}: *distance to Expressway at site* (s_i);
lat: *latitude of site* (s_i);
$long$: *longitude of site* (s_i);

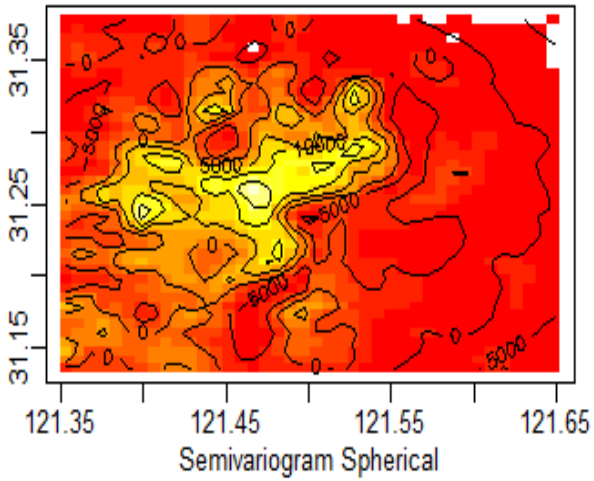
Selection based on AIC

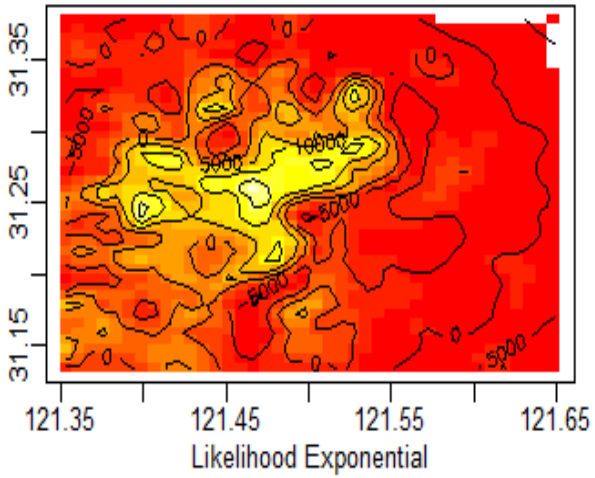
```
Parametric coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.005e-04  4.551e-06  44.057  < 2e-16 ***
D2Metro         6.947e-09  4.602e-09   1.509 0.131263
D2Road         -4.564e-09  1.777e-09  -2.569 0.010246 *
I(D2Metro^2)    7.422e-12  1.913e-12   3.880 0.000106 ***
D2Metro:D2Road  5.605e-12  2.441e-12   2.297 0.021692 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

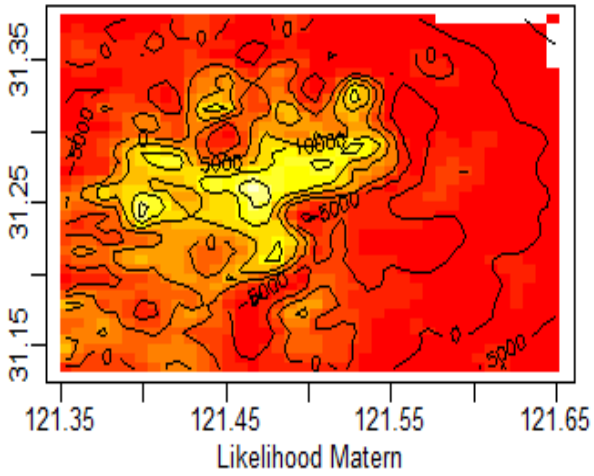
CONTENTS

PREDICTION





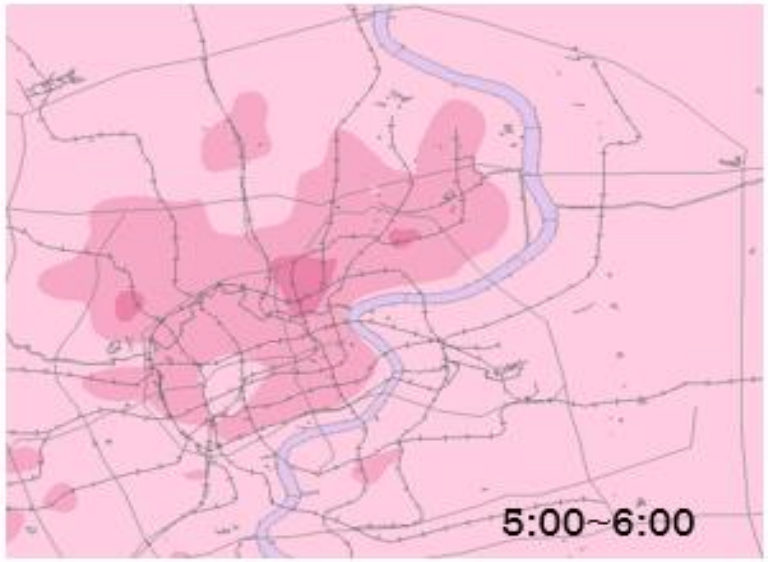


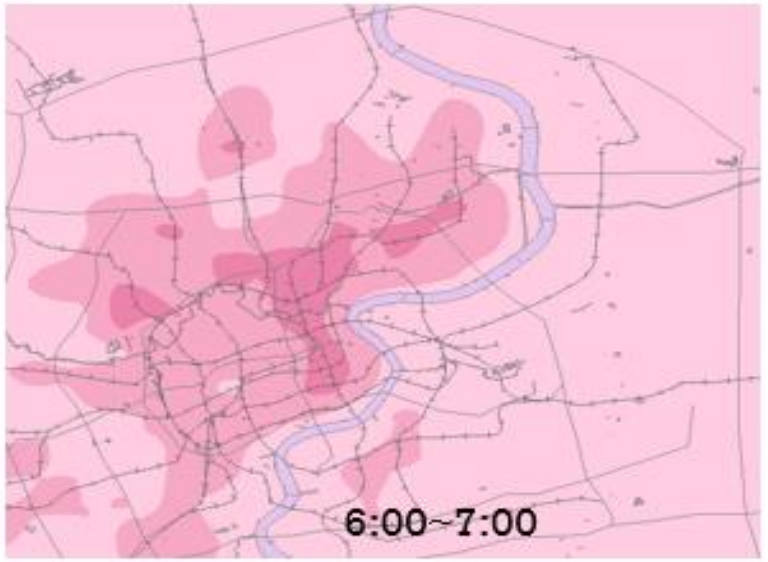


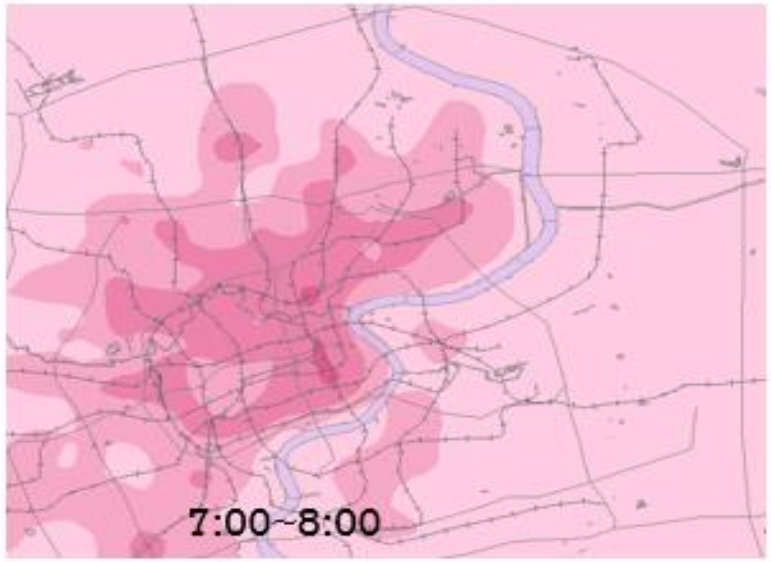
SSE for predict models:

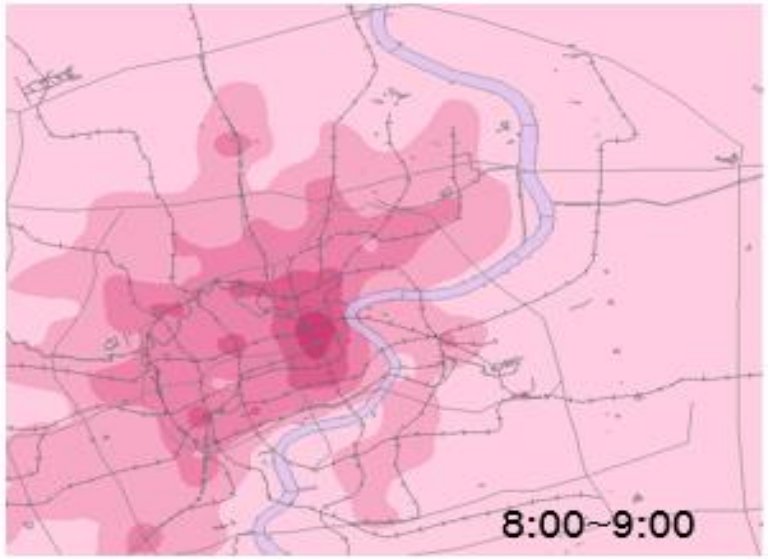
	Spherical	Exponential	Matern
5	483239486	441347900	441347900
6	640487307	585196252	585196252
7	1018205620	917097306	917097306
8	1172933529	1074494482	1074494482
9	1071913221	980410306	980410306

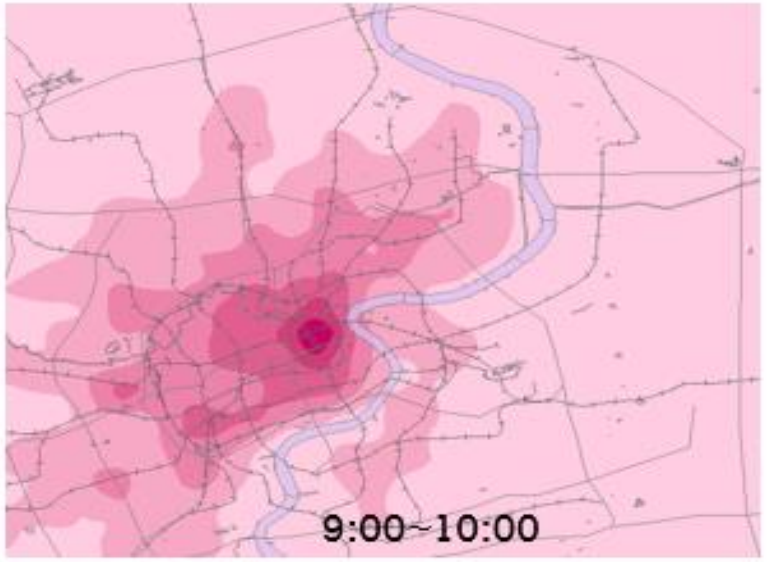
PREDICTION











CONTENTS

CONCLUSION

- More and more people moving toward CBD
- Population density variation increases by time
- The closer to Metro or Expressway, the more people
- Movement along transportation corridor

Spatial pattern of population movement during morning peak hours

Thank You!

STAT 5544 Final Report

Wenyu Gao, Danni Lu

Dec.1st, 2016