Street Crime Forcasting

Liuyi Hu

Department of Statistics, NCSU

Introduction

Police forces have significant interest in being able to predict regions in which crimes are likely to occur so that preventive measures may be employed in both the short-and long-term. This project aims to model both the temporal and spatial dependencies often exhibited by street crimes in order to make such predictions.

Data

- **Street crime records**
  The street crime records come from calls-for-service (CFS) records provided by the Portland Police Bureau (PPB) for the period of March 1, 2012 through August 31, 2015. These records contain the coordinates as well as date of the street crime.
- **Areal units**
  I divide the Portland Police District into 484 rectangles called "cells" and then aggregate the crime locations within the same cell together. Figure 1 shows the average monthly street crime counts over the cells.
- **Demographic covariates**
  The demographic information comes from Geographic Information Services (GIS). I consider the following demographic covariates: (1) **population density** (2) **unemployment rate** and (3) **white proportion**.

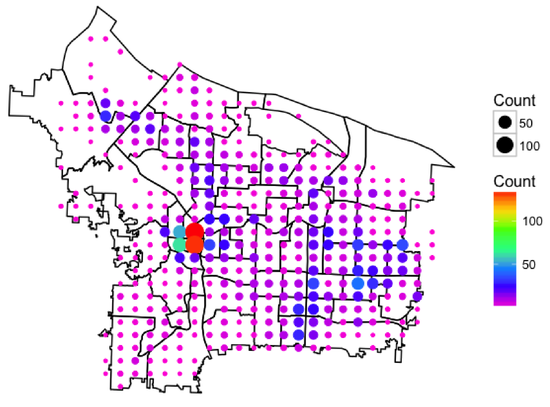

Figure 1: Average monthly street crime counts over cells.

Methods

- **Model 1**
  The temporal variation of crimes typically follows patterns familiar in time series analysis and Figure 1 implies that neighboring areas share some crime dynamics. Therefore I consider the following spatiotemporal model:

$$\log\left(y_t(s)+1\right) \mid \mu_t(s) = \mu_t(s) + \boldsymbol{x}_t^T(s)\boldsymbol{\beta} + \epsilon_t(s),$$

where

- $y_t(s)$ is the total street crime during month t and in cell s
- $\epsilon_t(s)$ is iid Normal $(0, \sigma_e^2)$
- $\mu_1(s)$ is a GP with mean 0, variance σ_s^2 and Matern correlation with range ψ and smoothness ν
- $\mu_t(s) \mid \mu_{t-1}(s)$ is a GP with mean $\rho\mu_{t-1}(s)$, variances $\sigma_s^2(1-\rho^2)$ and Matern correlation with the same parameters as μ_1

- **Model 2**
  Crime rates often vary seasonally, with a higher rate during warmer months of the year. Figure 2 shows the monthly counts of street crime from March 2012 to July 2015 for 4 selected cells. We can see that in all three years, there is a peak during summer time. So to capture the seasonal effect of crime, Model 2 includes the indicator of month as covariates besides the demographic covariates.

$$\log\left(y_t(s)+1\right) \mid \mu_t(s) = \mu_t(s) + \boldsymbol{x}_t^T(s)\boldsymbol{\beta} + \boldsymbol{z}_t(s)^T\boldsymbol{\gamma} + \epsilon_t(s)$$

where $\boldsymbol{z}_t(s)$ is the indicator vector for the month.

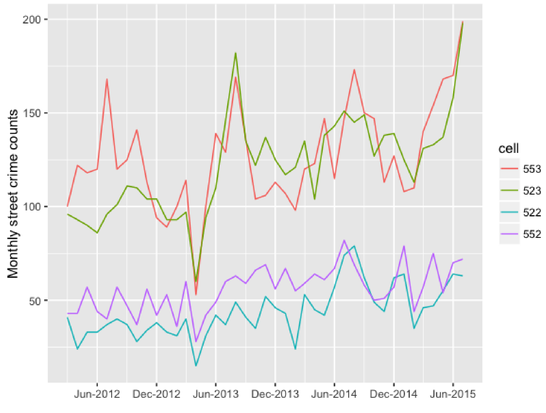

Figure 2: Monthly counts of street crime from March 2012 to July 2015.

Model Comparison

I use all data prior to August 2015 as training data and use August 2015 as test data. Table 1 summarizes the performance of two models on the test data. We can see the Model 2 outperforms Model 1 in terms of coverage and mean squared error.

Model	Coverage	Mean squared error
1	96.9%	0.182
2	97.2%	0.179

Table 1: Model comparison.

Results

Table 2 summarizes the parameter estimation using Model 2. Figure 3 shows the box-plot of posterior sample for month estimate (month 12 is the reference group). This implies that there is a seasonal effect and that the summer time tends to have more street crimes than winter time.

Parameter	Estimate (Std.Error)
σ_e (Nugget SD)	0.419 (0.003)
σ_s (Partial sill (SD))	0.945 (0.035)
ρ (AR coefficient)	0.999 (0.000)
ψ (Matern range)	2.573 (2.282)
ν (Matern smoothness)	0.926 (1.718)
population density	-0.029 (0.018)
unemployment rate	0.186 (0.015)
white proportion	-0.271 (0.063)

Table 2: Paramter estimate for Model 2. Standard errors are given in the parenthesis.

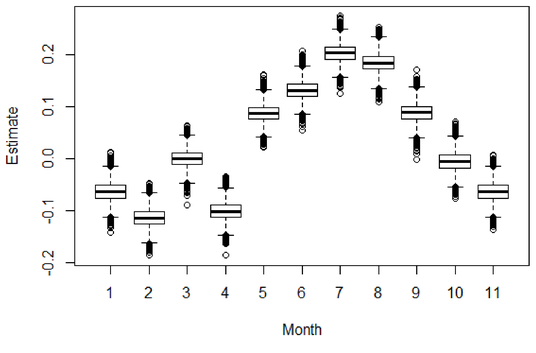

Figure 3: Box-plot of the posterior sample for month estimate.

- **Prediction Accuracy Index (PAI)**

$$\text{PAI} = \frac{398/3349}{2/484} = 28.8.$$

Figure 4 and 5 are the maps of predicted and observed values of street crimes counts during August, 2015 in each cell after log transformation.

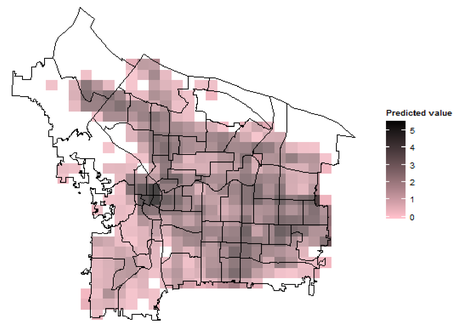

Figure 4: Predicted values for August 2015 after the log transformation.

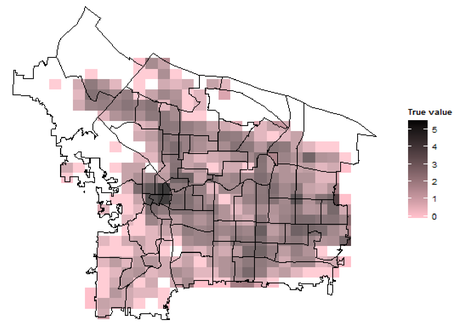

Figure 5: True values for August 2015 after the log transformation.

Limitations and Future Works

- For the response data, I took log-Gaussian transformation. Instead, we can also model the count directly as a Poisson process and assume the Poisson rate has a seasonal effect.
- I pre-determined the size of the cells in my analysis. To find the optimal size of the cells that maximizes the PAI score, we can use cross-validation.