Spatial analysis of Dinner Restaurants in Bangalore

Suhas Shastry

March 21, 2019

Introduction

Background

This project is a spatial point process analysis of all the dinner restaurant in Bangalore. Bangalore political map in `geoJSON` format was obtained from **Data{Meet}** community. This data has boundary information of 198 wards in Bangalore along with population and area of each ward.

Using this ward information, all the dinner restaurants in these wards were collected from **Foursquare** using their APIs.

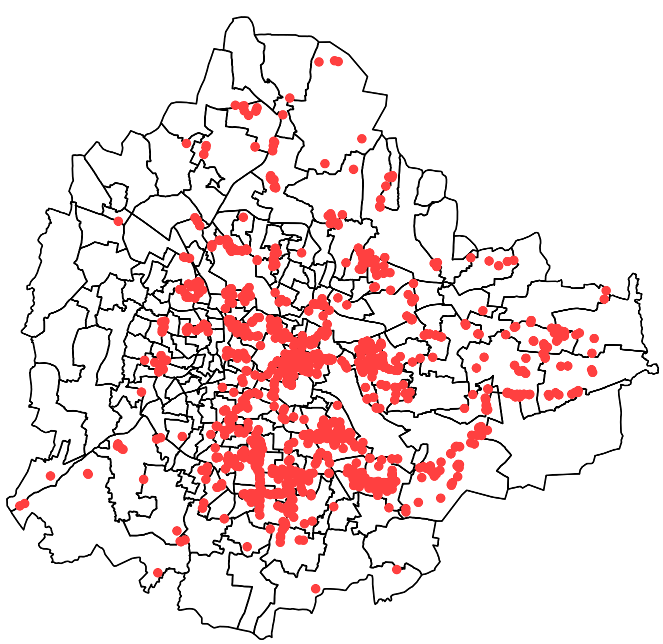
Data Modification

Restaurants details obtained from Foursquare contained duplicates. Once they were removed, 1405 different dinner restaurants' details was retained. These restaurants belonged to different wards but the exact information to which ward it belongs to is unknown. This information was obtained using `point.in.polygon` function in R. Finally proximity matrix for 198 wards is required for spatial analysis. Proximity matrix was prepared using First Order Queen Contiguity. FOQ contiguity defines a neighbor when at least one point on the boundary of one polygon is shared with at least one point of its neighbor (common border or corner).

Analysis

Before we do any analysis, we plot all 1405 restaurants in Bangalore in Figure 1.

Figure 1. Point plot of dinner restruants in Bangalore



Restaurants are concentrated in central and southern parts of Bangalore. There are many wards in Eastern Bangalore without a single dinner restaurant within them. Factors that might affect the number of restaurants in a ward are location of the ward, population in that ward, area of that ward. Based on these factors, we fit a simple linear regression model of the form, $y = \mathbf{X}^T\beta + \epsilon$. This model has R^2 value of 0.116 and AIC of 1566.598 which reflects the poor explanation by a linear model. We even tried Poisson regression model of the form,

$$\eta_i = \log \lambda_i = \mathbf{x}_i^T \beta + \epsilon_i$$

where λ_i is the rate parameter in ward i for $i = 1, \ldots, 198$ and ϵ_i is the error. This performs worse than linear model with AIC 2877.532. Hence we explored spatial analysis for the data under consideration.

Spatial Poisson regression model

Let y_i be the number of restaurants in each ward, then $y_i \sim Poisson(\lambda_i)$. We explored spatial Poisson regression model of the form,

$$\eta_i = \log \lambda_i = \mathbf{x}_i^T \beta + \psi_i + \epsilon_i$$

Here λ_i is the rate parameter of each ward, ψ_i is the spatial random effect and ϵ_i is the spatial effect. We consider a hierarchical model for the above regression model. That is, $\beta \sim N\big(\mu, \kappa_1^{-1}(X^TX)^{-1}\big)$, $\psi \sim N\big(0, \kappa_2^{-1}Q^{-}\big)$ and $\epsilon \sim N\big(0, \kappa_3^{-1}I\big)$. We assume that newly introduced κ_1, κ_2 and κ_3 are i.i.d $Gamma(a, b)$ random variables. We know 198 y_is and $\mathbf{X}$ matrix, but the unknown parameters are 4 β's, 198 ψ_i's, 198 η_i's and κ_1, κ_2 and κ_3. Totally there 403 unknown parameters that are to be estimated with known 198 values. We proceed to this estimation using Gibbs sampling.

Gibbs sampling

Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. In our problem, we know values of y and $\mathbf{X}$. We want to estimate values of η, β, ψ, κ_1, κ_2 and κ_3. We use the Bayesian approach to estimate the joint distribution of all these parameters given data.

$$\pi(\eta, \beta, \psi, \kappa_1, \kappa_2, \kappa_3 | y, \mathbf{X})$$

By law of total probability, this joint distribution is the product of all marginal distributions.

$$\pi(\eta, \beta, \psi, \kappa_1, \kappa_2, \kappa_3 | y, \mathbf{X}) = \pi(\eta | \beta, \psi, \kappa_1, \kappa_2, \kappa_3, y, \mathbf{X}) \pi(\beta | \kappa_1, y, \mathbf{X}) \pi(\psi | \kappa_2, y, \mathbf{X}) \pi(\kappa_1) \pi(\kappa_2) \pi(\kappa_3)$$

These individual marginal distributions can be simulated from the known distributions using Gibbs sampling.

Simulation

Joint distribution was simulated with initial estimates for $a = 1$ and $b = 0.005$. $\beta^{(0)}$ was taken from linear regression model. 5000 simulations were run 1000 burn in period.

Figure 2. Simulated values of Kappa

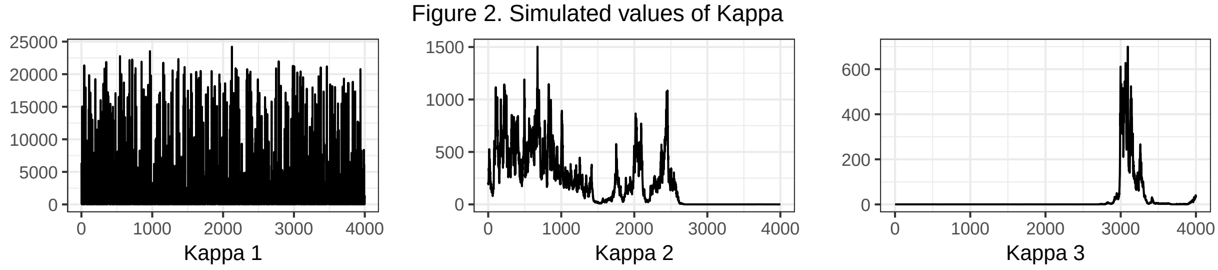


Figure 2. indicates the simulated values of κ's. Almost all three of them stabilize in the long run.

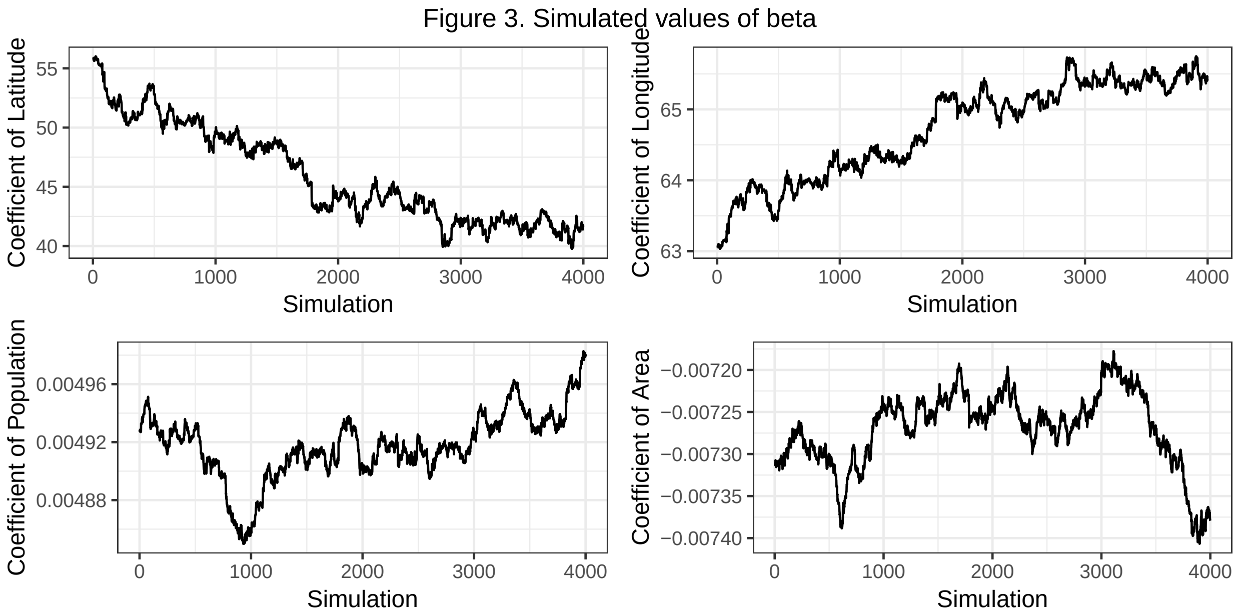


Figure 3. indicates the simulated values of 4 different β's. There is a trend in almost all four of them. This could be overcome by simulating for long time. But for now, we are considering average of these values for interpretation.

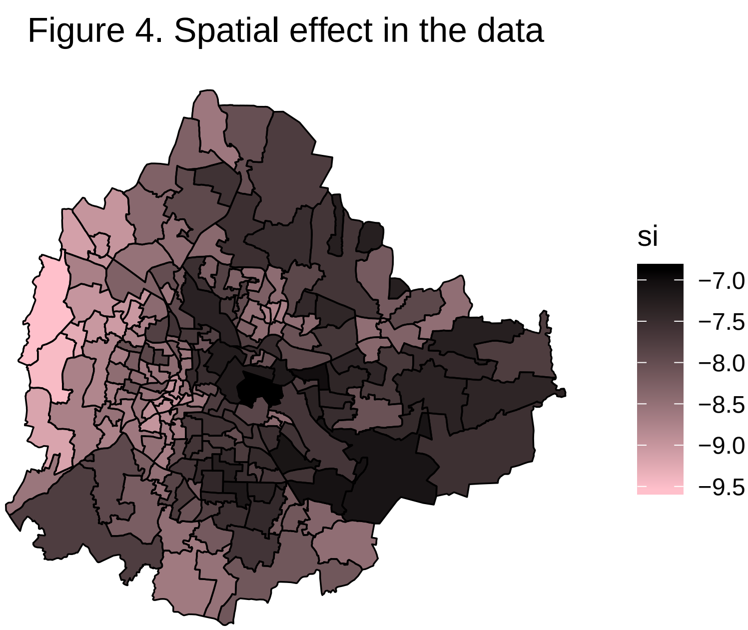


Figure 4 indicates the spatial effect in the data. Spatial effect is high in central and southern parts of Bangalore which is in correlation with density of restaurants. Eastern parts of Bangalore have very low spatial effect. Figure 5. indicates the residual effect. Residual is close to zero in all most all the wards. Residual is bit high in the wards which have very high or very low restaurants compared to their neighboring wards.

Figure 5. Residual effect

Figure 6. Rate parameter eta

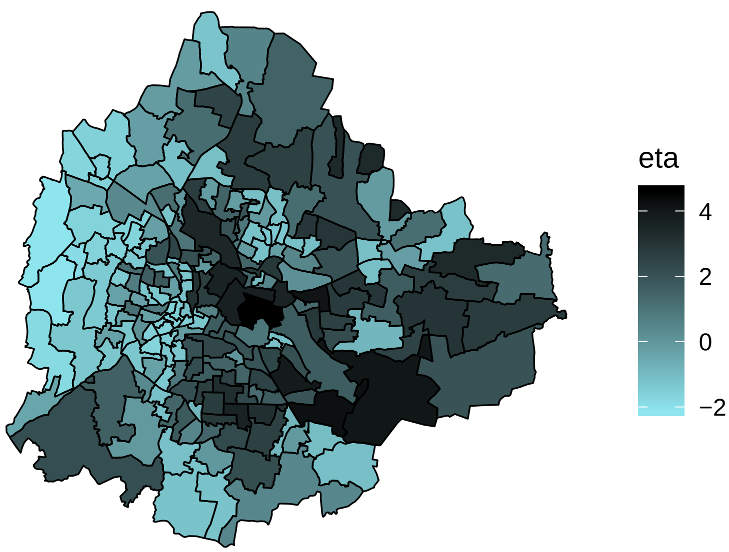


Figure 6. indicates the estimated log of rate parameters for different wards. As expected, rates are very low in eastern wards and very high in central and southern wards.

Conclusion

We considered details of 1405 restaurants in Bangalore and divided it into 198 wards. Each restaurant was considered as a point process. Initially, linear regression model and Poisson regression model were fit. As these models failed to explain the variation, we fit spatial Poisson regression model. Because we had more unknown paramters, we used Gibbs sampling to estimate all these paramters. Based on these simulated estimates we plotted spatial, residual effects for each ward and also we displayed log of rate parameter for each of the ward.