STA 644 Project Report

Yaqian Cheng, Yulin Lei, Mengrun Li, Leonardo Shu

May 1, 2017

1 Introduction

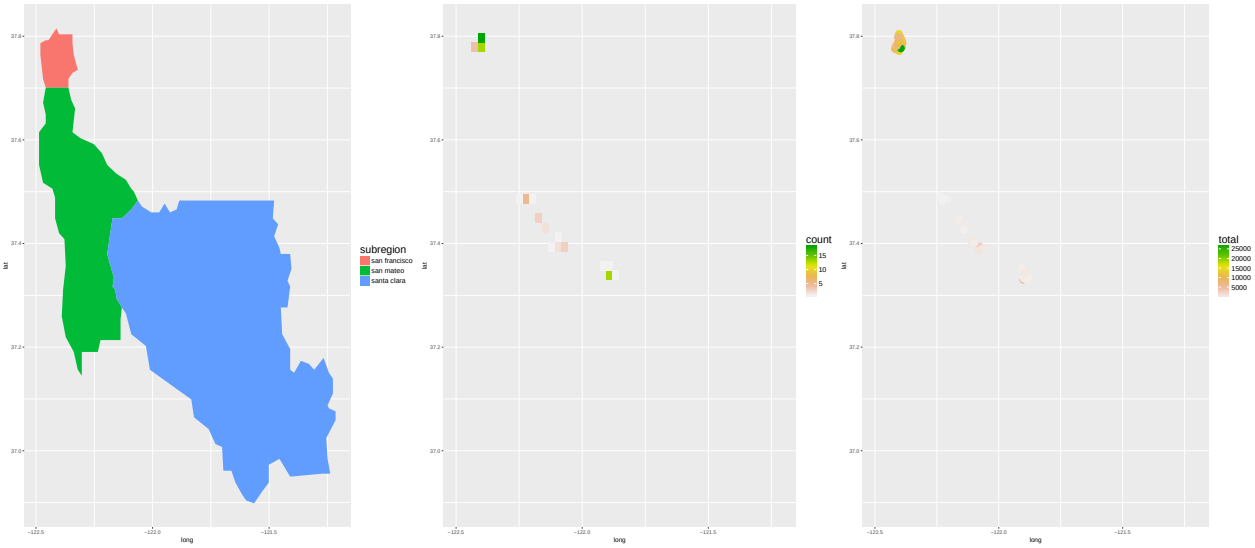
Our aim for this project was to explore the models concerning point reference data and observe how they performed on a scenario of interest to us. Specifically, we wanted to see how Gaussian Process Models and Thin Plate Splines could help us make use of bike trip data in the Bay Area by predicting the locations of where users would most likely start a trip from. In doing so we would be able to see which bike stations are the most popular (in demand) and whether there are other areas they could expand to in order to capture more potential customers.

2 Data Description

We downloaded the data from http://www.bayareabikeshare.com/open-data. There are two datasets that we mainly used recorded from 2014/09/01 to 2015/08/31 including station information(station ID, name, latitude, longitude, dockcount, city), trip information(time, start terminal, end terminal, duration in second). There are 70 bike stations located in 5 different Bay Area cities, which are San Francisco, Palo Alto, Redwood City, Mountain View and San Jose.

The target response variable we defined is the number of trips starting from a specific bike station. We combined the two datasets and aggregated into yearly, day of week and hour of day level of each station, respectively.

2.1 Exploratory Data Analysis



The figure on the left above shows the three counties(San Francisco, San Mateo and Santa Clara) where the

5 cities are located. The one in the middle is a heatmap of count of stations. As can be seen, there are more bike stations in San Francisco and San Jose than they are in the other three cities located in San Mateo county. The figure on the right shows the total number of trips starting from the given station through the whole year. We can see that most of trips are in San Francisco, which makes sense. Because San Francisco has higher population density than the other four cities and it also has more bike stations. According to these findings, we will fit two spatial models in the next section to capture and predict the bike trip pattern of three counties.

3 Model

3.1 Gaussian Process (GP)

We used a Gaussian Process model to fit the data and assumed an exponential covariance structure.

$$\boldsymbol{y} \sim (\mu, \Sigma)$$
$$\{\Sigma\}_{ij} = \sigma^2 \exp(-r||s_i - s_j||)\sigma_n^2 \mathbf{1} i = j$$

Where $\boldsymbol{y}$ is the total/average number of trips starting from selected coordinates.

To fit Gaussian Process model, we used `spLM` function from the package `spBayes`, where predictors are longitudes and latitudes of those 70 stations and response variables are the total/average number of trips starting from those stations. We set `starting` parameter values according to the variogram, use default values for `tuning`, and choose `prior` parameters according to the `starting` parameter values.

A raster is a spatial (geographic) data structure that divides a region into rectangles called 'cells' (or 'pixels') that can store one or more values for each of these cells. Such a data structure is also referred to as a 'grid'. We generated the raster based on the boundary of bay area. After we got predicted values from fitted model with coordinates for prediction `pred_coords`, we could fill in the raster and plot the result.

3.2 Thin Plate Splines (TPS)

Observed data: (x_i, y_i, z_i), where we wish to predict the number of trips z_i given longitude x_i and latitude y_i for all i

The smoothing spline model in two dimensions:

$$\underset{f(x,y)}{\arg\min} \quad \sum_{i=1}^{n}(z_i - f(x_i, y_i))^2 + \lambda \int\int \left(\frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial x\, \partial y} + \frac{\partial^2 f}{\partial y^2} \right) dx\, dy$$

Solution:

$$f(x, y) = \sum_{i=1}^{n} w_i \; d(x_i, y_i)^2 \log d(x_i, y_i).$$

To fit TPS model, we used `Tps` function from the package `fields`, where predictors are longitudes and latitudes of those 70 stations and response variables are the total/average number of trips starting from those stations. Then we used prediction coordinates, `pred_coords`, and the TPS model as input to predict the average number of trips, `trip_pred`.

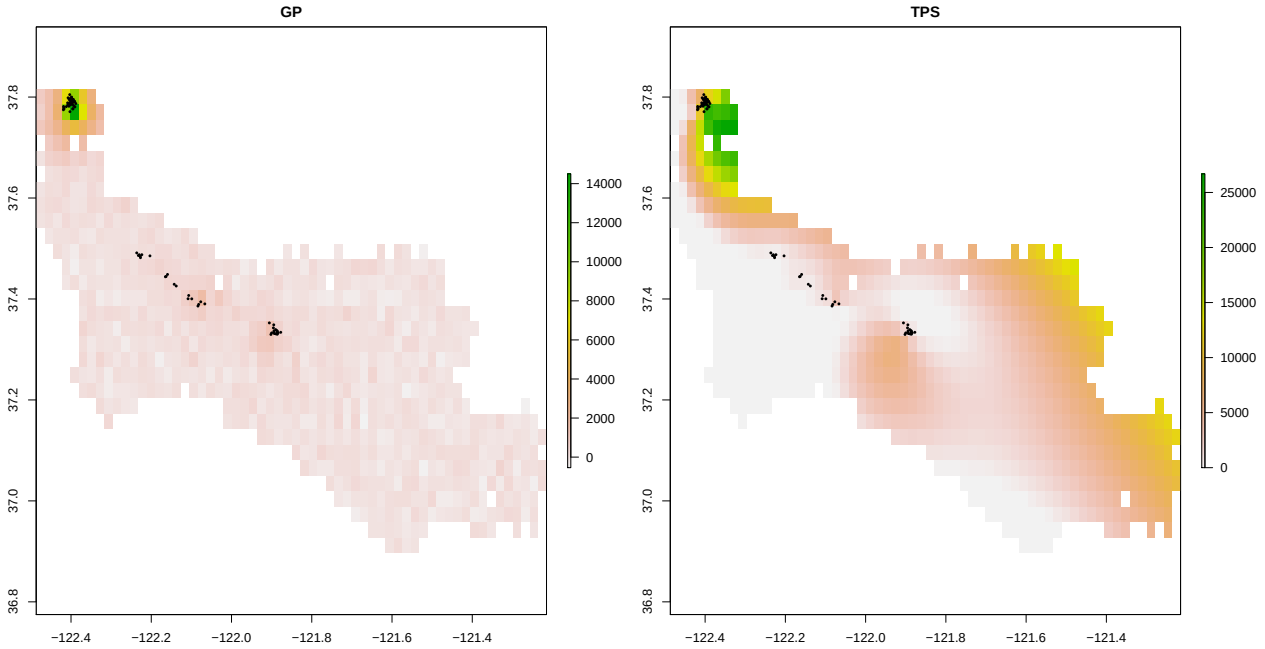
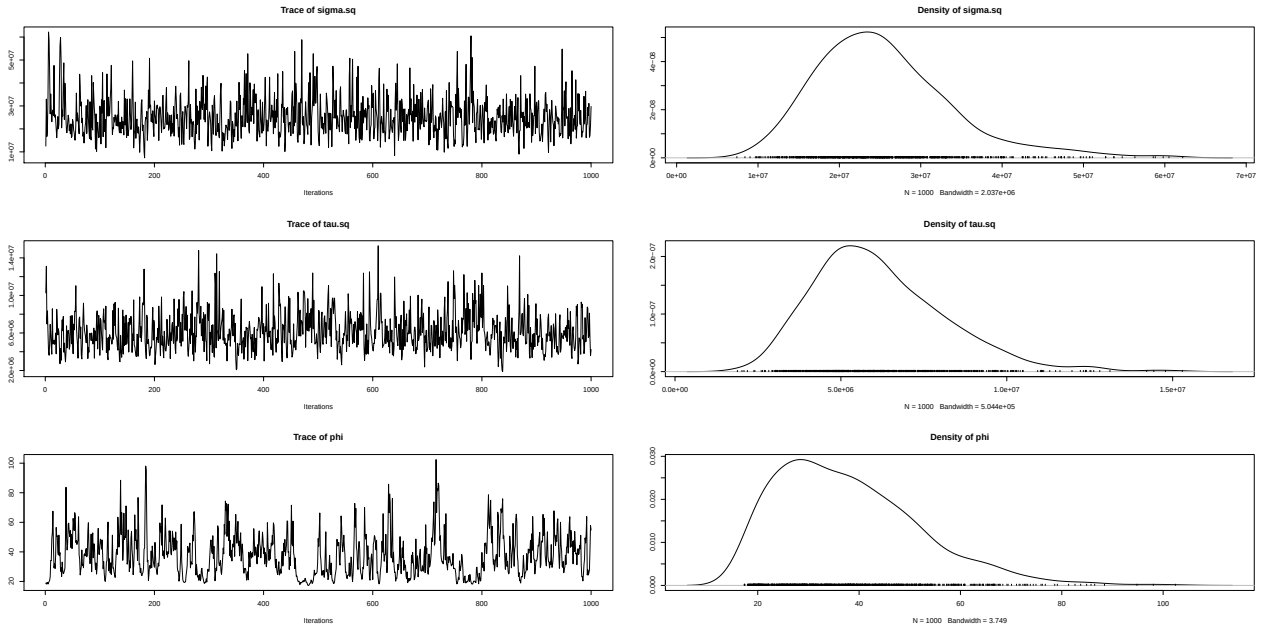
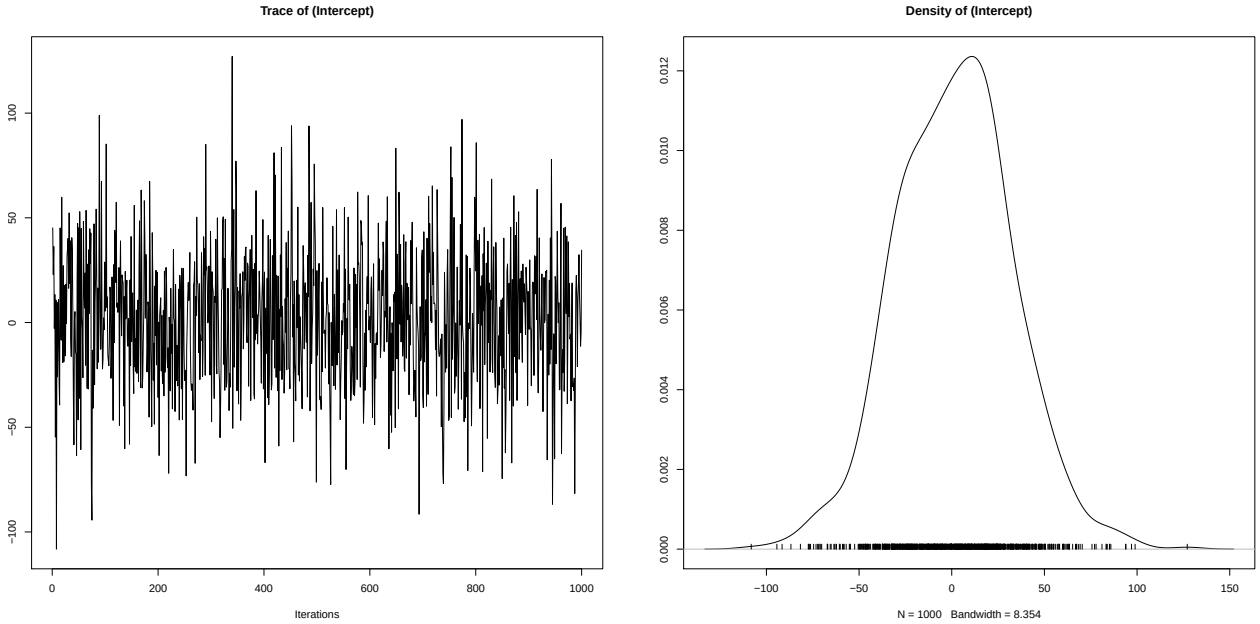


Figure 1: Overall Prediction Mapf for GP and TPS

4 Model Fits and Interpretation





4.1 OVERALL

4.1.1 GP

Correctly determines the popular demand around San Francisco and the low demands amongst the stations in Palo Alto and San Jose. Everywehere else where there is not much training data it predicts with the mean number of trips overall. We also see from the above traceplots that all parameters are converging well so the problems are how we choose our tuning and prior parameters.

4.1.2 TPS

This model provides much more intuition with the smoothing it performs. It tells us that The southern areas of the San Francisco is where most of the demand is and it could bode well for the company to add more stations there. Palo Alto and San Jose don't see many trips a priori so it makes sense that the areas around these stations do not see alot of trips. In fact, this model predicts that alot of areas such as south of San Mateo and northeast of San Jose have no expectred trips and this is something we should expect given that the smoothing should not that far from the points we do have data on.

4.2 DAY OF WEEK

4.2.1 GP

Different, but constant patterns across weekdays and weekends. The model does not do well in predicting areas far away from our known stations so it basically gives a mean value to these. In San Francisco, however, it seems to ascertain that most of the trips are happening there. There's also a notable difference in scales and we see much less trips on weekends. This is counter-intuitive since we first thought these bikes were used by tourists as a leisure activity but perhaps they are being used by locals to commute to work.

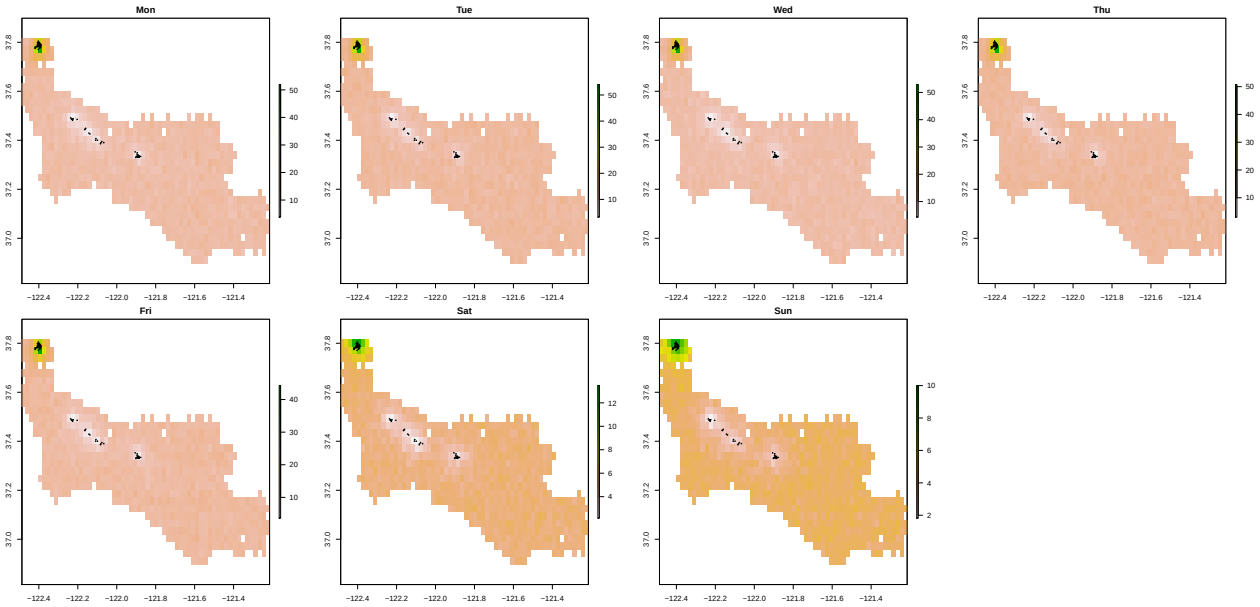


Figure 2: Day of Week (Mon-Sun) Prediction Map for GP

4.2.2 TPS

Once again we see very distinct patterns between weekdays and weekends. Unlike before, there are many patches were the estimated number of trips will be 0 such as south of San Mateo during the week. The areas around San Francisco are still the most popular to start trips but we still predict some trips around Palo Alto and most of Santa Clara. We think these are the places where people are working the most. This supports the idea that locals use these bikes to work since these areas are predicted to have 0 trips during the weekend. Also during weekends there again is less trips on average so using them for leisure is likely.

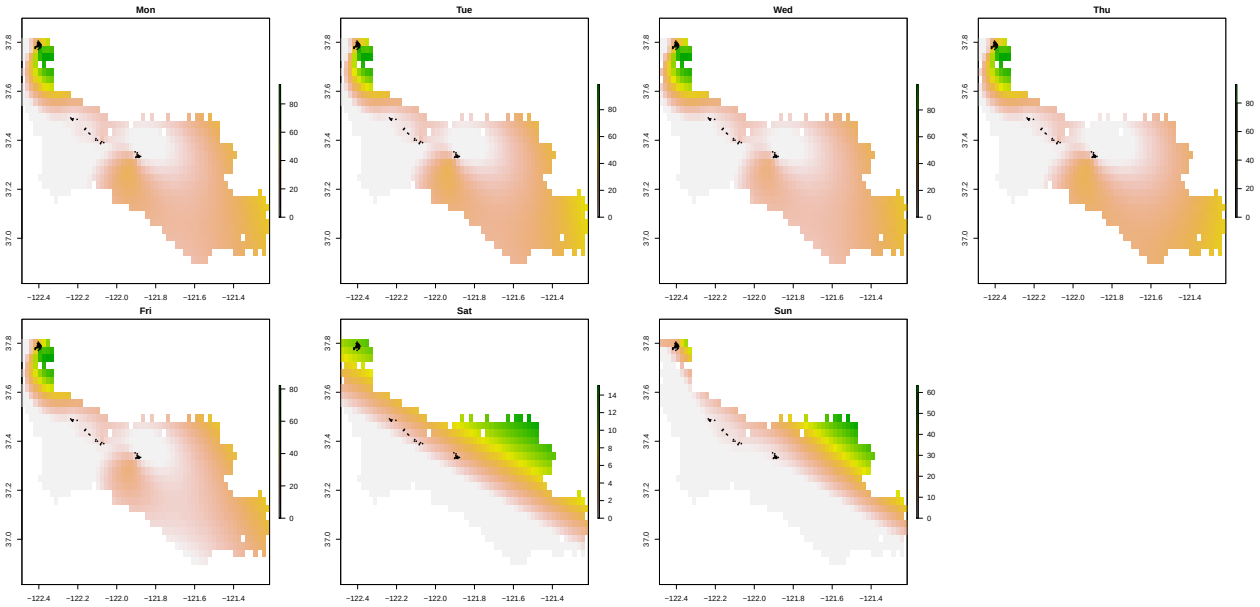


Figure 3: Day of Week (Mon-Sun) Prediction Map for TPS

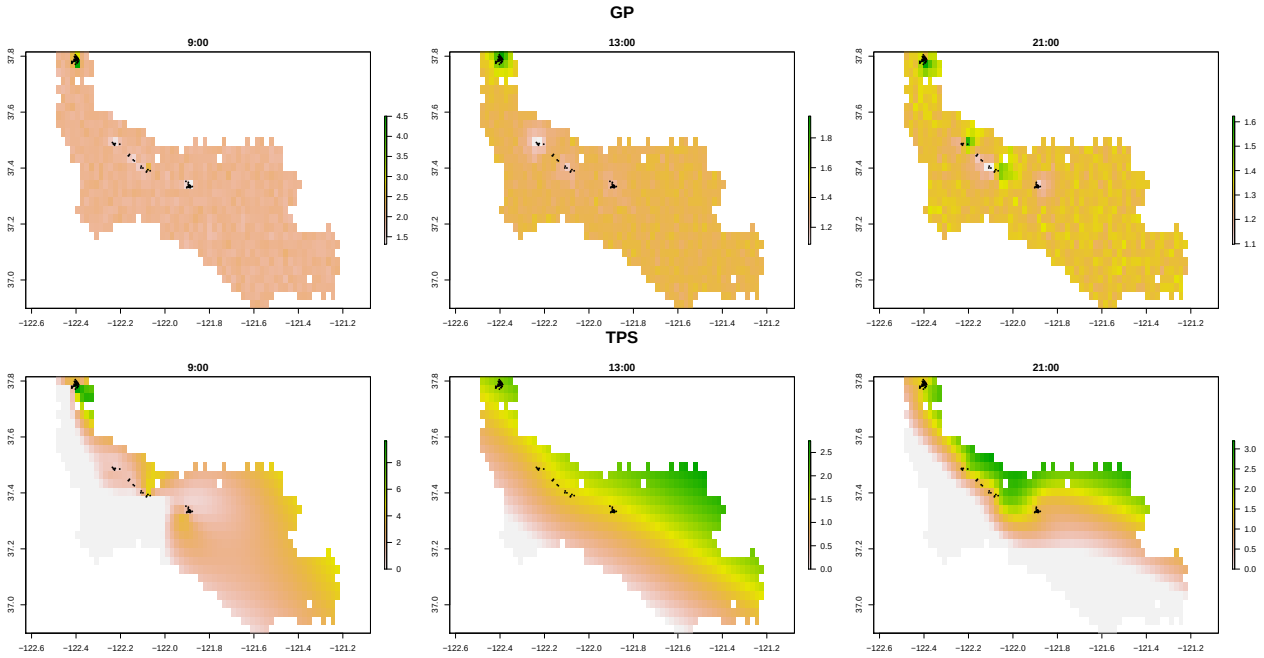
4.3 HOUR OF DAY

4.3.1 GP

Mornings are similar to day of week where the SF stations are popular and the rest of the map is predicted evenly. Same pattern early afternoon but with less trips overall (less than 2). Nighttime has the same pattern but now there are more trips all across the bay area. Which is more than we have seen before so night time seems to be quite an active period where leisure is now used.

4.3.2 TPS

Commuting time sees south of San Mateo to be blank as before reinforcing the idea that most people that commute to work do not have jobs there. Early afternoon looks like the overall pattern and less trips overall in scale. Same pattern as in early afternoon with overall more trips around the area which means this time of night is pretty active for bike riding.



5 Conclusion

All in all even though are predictions are quite raw and we forced many assumptions on the data so we could fit these models, we believe most of the interpretations we did manage to get could be valuable to the Bay Area Bike Share company since they could re-evaluate if its necessary to keep many stations in Redwood City or Palo Alto which see little demand during the day and can instead focus their resources on keeping all the SF stations supplied giventhat most of their businesss is there. For further improvements we would like to try these models with data sets that cover more ground and are not as limited to particular locations. We also would like to try a wider range of tuning parameters and prior values so the GP model acts more as expected and does not breakdown as often.