1 Introduction

The goal of the project is to fit a model that can predict incoming flows into a region in a city to aid the process of city planning. The project takes advantage of the recent abundance of massive passive phone calling data to infer the movement of the population in a city and explores the potentials of further developing city and urban planning tools using probabilistic models. I used a data driven approach to estimate the existing flows in the city from the raw data [5]. The output of the algorithm implemented is a set of l_{ij} representing a validated estimates of the number of people moving from region i to region j in a typical day. I used the validated estimates to develop a probabilistic model that can predict those flows given the flows of other regions of similar attractions. Then, I compare the results of my approach to the gravity model which is a common method to estimate trips in the domain of transportation engineering.

2 Methods

I first used the algorithm developed in my lab to estimate the flows in the city of Riyadh in Saudi Arabia. The method uses raw Call Detail Records from a telecom company provider which is the Saudi Telecom Company (STC) and outputs validated flows l_{ij} [5]. Given people's flow data in the city, I went through various attempts at modeling the problem starting with a Dirichlet-Multinominal model where the Dirichelet distribution has a simplex of the types of places (attractions) and the distribution models the variability of regions in terms of the places that are in them, the multinomial would then model the flows. The model didn't fit the problem quite well where I then moved to an implementation of a Gaussian process for the problem. My initial attempts discussed earlier in the progress report suffered accuracy issues and computational challenges that didn't end as I wished. Then I moved on to developing a spatial Gaussian process model that learns the pattern of inflow of those regions that have similar attraction profile and use the model for prediction. The intuition behind the modeling approach is that inflows of people are usually driven by the places in a destination region and similar regions exhibit similar spatial inflow signatures [1, 6]. For example, to model the inflow for a university in a city, we fit a model on the other existing universities in the city and use the model to predict the inflow. Therefore, we utilize our prior knowledge about the visitors of universities where they usually come from similar locations.

2.1 Gaussian Process Model (GP)

The Gaussian Process model parametrizes the incoming flows to regions with similar attraction profiles by their geographical coordinates. We define a set k as the set of regions with similar places of interest. In the universities example, the set k represents the set of regions having universities in them. A sample of the data is included in table 1. Each row in the data has the form $[lat_i, lon_i, l_{ik}]$ meaning that we have an inflow of l_{ik} people from the geographical coordinate $[lat_i, lon_i]$ where i is the index of the region that is the source of the flow and k is an index of regions with similar attraction places.

The first step towards this problem is to define a Gaussian process for the input data $\{x_1, x_2\} \subset X$ where x_i represents a row of the parameters discussed above that is $[lat_i, lon_i]$ and l_{ik} is the flow from that point. The Gaussian process is parameterized by a mean function $m(x)$ and a covariance function or kernel $k(x, x')$ where we get a finite set of functions equal to the number of points we

i	lon_i	lat_i	l_{ik}
1	46.5766	24.7173	45
2	46.5194	24.7461	68
3	46.5920	24.7166	6

Table 1: sample of inflow data for regions set k from region i

have. The Gaussian process is then given by

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{1}$$

Where the mean function and kernel function are given by

$$m(x) = 0, \qquad k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l}\right)$$

To implement the model, I used the `Matlab GPML toolbox` developed by Carl Edward Rasmussen and Hannes Nickisch [4] which is an implementation of the topics covered in their book [3]. The method allows for various choices of mean functions, covariance functions, likelihood functions and inference methods. For the purpose of this project, I used a Gaussian likelihood and used exact inference for the parameters. Exact inference is computationally feasible in this model as the number of points is $\approx$ 1500 which is the number of regions in the city of Riyadh shown in Figure 1. My choice of the hyper parameters σ and l was based on experimentation of the output of the model where I found $\sigma = 1, l = 0.01$ to be performing reasonably well.

3 Results

This section includes the results of the implementation of the model on the city of Riyadh in Saudi Arabia. I will be estimating the flows into King Saud University shown in red in the figure below. To do that, I will develop a GP for spatial inflow to the regions in the set k defined as the regions with universities in them and shown in the green color in the figure below.

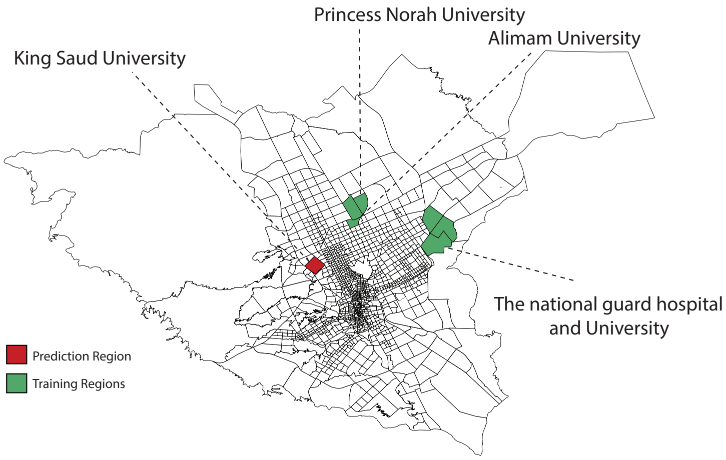

Figure 1: Training regions versus prediction region, training regions represent the set k

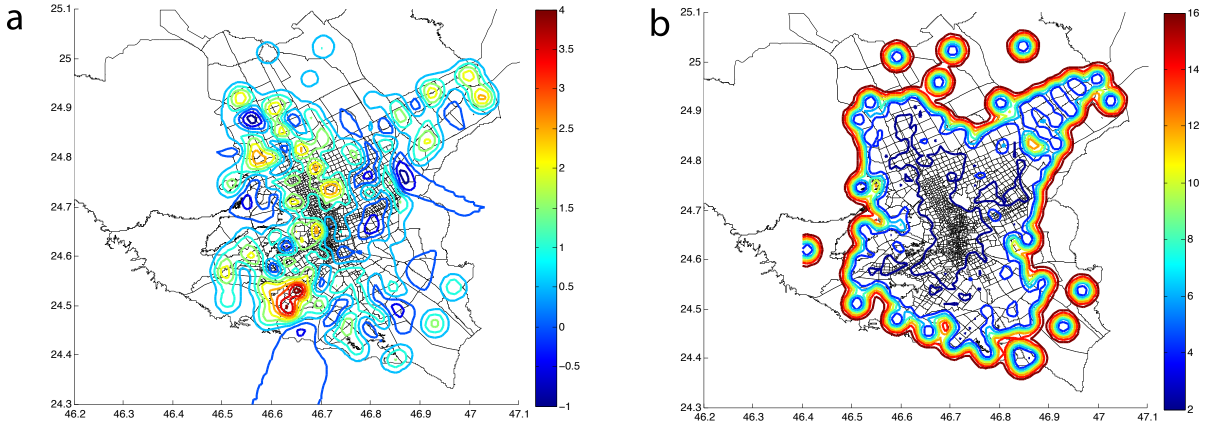

Figure 2: The figure shows (a) log of predictive mean and (b) predictive variance for the inflow to regions in the set k

Figure 2-a shows the mean predictive after training the GP on the inflows to regions in the set k. The model captures the main sources of inflows to universities. We can see that there is a major inflow to universities from the south-western region of the city which is known to be highly residential. The significant inflows in general significantly overlap with highly residential regions. Figure 2-b shows the variance around the mean predictive of of inflow. The variance is relatively low inside the boundaries of the city where we have inflow data and is highest outside the city where no inflow data is available. I will use the value of the mean predictive when predicting inflows to King Saud University.

3.1 Evaluation

The metrics I used in quantifying the accuracy of the model are the Root Mean Square Error (RMSE) and the Mean Error (ME) given by

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(L_{ij} - l_{ij})^2} \quad , \qquad \mathrm{ME}_j = \frac{1}{n}\sum_{i=1}^{n}|L_{ij} - l_{ij}|$$

where L_{ij} is the predicted flow to location j from location i and l_{ij} is the actual flow. The ME is more interpretable for me to evaluate the model in terms of the average error of flow quantities but I also used RMSE to evaluate how far the predicted flows are from the actual values.

3.2 Baseline model

For the purpose of evaluating the performance of the model compared to existing methods, I compare the results of the model to the gravity model as the baseline model used in the domain of transportation engineering [2]. The model is given by

$$L_{ij} = \frac{O_i \, T_j}{d_{ij}^{\alpha}}$$

where O_i is the total outflow from a location i and T_j is the total inflow into location j and d_{ij} is the street distance between the i and j regions and α is a calibration parameter to be estimated.

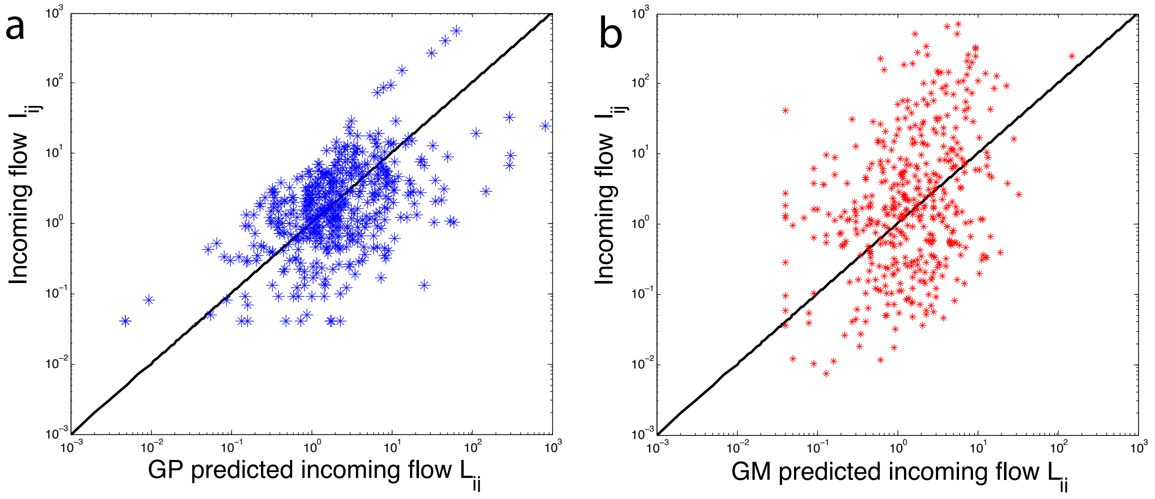

Figure 3: Performance of (a) Gaussian process (GP) versus (b) gravity model (GM) for predicting the inflow to King Saud University

There are many versions of the gravity model, variations are always in the number of calibration parameters. This is a result of the lack of generalization of the model between cities which is a drawback of gravity models resulting in the disadvantage of overfitting the data. In our example, I chose a moderately complex version where I have one calibration parameter $\alpha = 4.8$ for the city of Riyadh.

Figure 3 shows plots of the predicted flows using the Gaussian process in (a) and using the gravity model in (b) versus the actual inflow values to King Saud University. The figure shows that the predicted inflows using the GP are closer to the $y = x$ line than that of the gravity model. Table 2 shows the performance of the Guassian process compared to the gravity model in terms of ME and RMSE where we find that GP has a ME of 10.5 people and a RMSE of 54.58 compared to that of the gravity model have a ME of 212.05 and a RMSE of 900.52.

method	ME	RMSE
Gravity model	212.05	900.52
Gaussian process	10.5	54.58

Table 2: ME and RMSE for GP and gravity model in predicting inflows to King Saudi University

4 Conclusion

The project proposes a probabilistic modeling approach for predicting the flows between regions in a city using phone calling data. The modeled GP estimates inflows more accurately than a gravity model fitted on the city of Riyadh. The initial results found in this report suggests that probabilistic models enabled by the abundance of phone data can sometimes provide better decision tools for urban planners than existing models in the literature of transportation engineering. I chose this project to experiment with the potentials of such models in city planning problems compared to traditional methods. I think it is worthwhile to extend the work to other cities and regions of other functionalities for further investigation of how GPs compare to gravity models.

References

[1] Alhazzani, M., Alhasoun, F., Alawad, Z., and Gonzalez, M. Urban attractors: Discovering patterns of regions attraction in cities. *Submitted to UbiComp: Ubiquitous Computing* (2016).

[2] Erlander, S., and Stewart, N. F. The gravity model in transportation analysis: theory and extensions. 18–19.

[3] Rasmussen, C. E. Gaussian processes for machine learning.

[4] Rasmussen, C. E., and Nickisch, H. Gaussian processes for machine learning (gpml) toolbox. *The Journal of Machine Learning Research 11* (2010), 3011–3015.

[5] Toole, J. L., Colak, S., Alhasoun, F., Evsukoff, A., and Gonzalez, M. C. The path most travelled: mining road usage patterns from massive call data. *arXiv preprint arXiv:1403.0636* (2014).

[6] Yuan, J., Zheng, Y., and Xie, X. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012), ACM, pp. 186–194.