ORIGINAL ARTICLE
Journal Section

Driving Miss Daisy: Bayesian Data Analysis of Thirty Million Uber Trips in NYC

David Arthur[1*] | Krishnan Raman[1*]

[1]Dept of Statistics, Purdue University

Correspondence
David Arthur
Email: arthur22@purdue.edu

Uber is a popular ridesharing service that connects millions of passengers with its drivers each day. Drivers are concerned with maximizing revenue over routes plied. We use Gaussian processes to model number of daily and hourly Uber trips at different locations in NYC. We also compute the posterior fare distribution for the optimal daily circuit. We compare fares between the optimal and random daily circuits to quantify wages lost by picking sub-optimal routes. We find that holidays are less busy times for Uber drivers compared to weekends, especially Saturdays. Areas in central NYC see the most Uber traffic. Mid-morning and early evening are also busier times. Picking a suboptimal route results in over $200 daily wage loss.

KEYWORDS
Gaussian Process, Spatio-Temporal Model, Posterior Distribution, Hierarchical Models, MCMC, Optimal Path, Confidence Distribution

1 | INTRODUCTION

Uber is a popular ridesharing service that connects millions of passengers with its drivers each day. Cabbies are primarily concerned with maximizing their daily take-home fare. Knowing the demand distribution i.e. what times of year, week, or day do most people hail an Uber, is vital to a cabbie. Additionally, it would be helpful to know locations that are busiest. Finally, an optimal daily route, one that makes the most money day after day (and is hopefully unknown to the vast unwashed masses out there not privy to the results of this paper) is the ultimate holy grail.

Cabbies spend most of their lives driving on the road. This here paper is a 'road map', that directs the Uber driver towards maximum revenue while they ferry their passengers from source to destination. A revenue GPS, if you will.

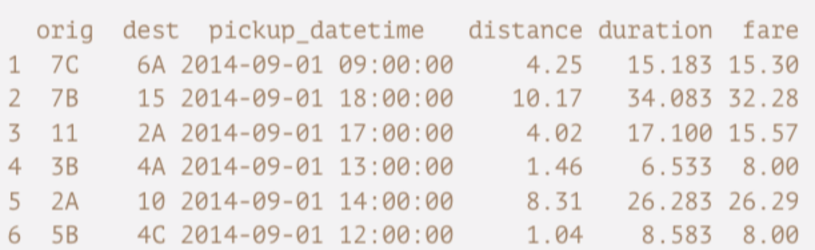
We explore the following questions:

1. What times of year and week are the busiest?
2. What times of day and what areas in NYC are the busiest?
3. What's an optimal daily route in NYC that makes the most money?
4. If the cabbie foolishly follows a suboptimal route, how much would they stand to lose every day?

This paper is organized as follows: First we discuss the data used in our analyses. We then introduce the Gaussian process which is a method upon which several of our models are built. The sections that follow will illustrate how Gaussian processes can be used to answer Questions 1 and 2. In these sections we will utilize the Gaussian processes to model the number of daily Uber trips over the course of a year as well as the number of hourly Uber trips across space and time. To address Question 3, we delve into how an optimal daily circuit is found. We introduce two Bayesian models, a simpler Gaussian Conjugate model and a more complex Hierarchical model, to compute the median fare obtained by following this optimal circuit. This process will help us answer Question 3. Finally, to answer Question 4, we will discuss confidence distributions and how they can be used to compare the posterior distributions for fare from the optimal and random circuits.

2 | DATA

A truly massive dataset comprising thirty million trips made by Uber drivers during the Sep 2014 - Sep 2015 timeframe was obtained from the The New York City Taxi and Limousine Commission. Despite anonymization, the dataset provides sufficient granular detail on time and place, fare and distance. A snapshot is shown below.



```
  orig  dest  pickup_datetime    distance duration  fare
1  7C     6A 2014-09-01 09:00:00     4.25   15.183 15.30
2  7B     15 2014-09-01 18:00:00    10.17   34.083 32.28
3  11     2A 2014-09-01 17:00:00     4.02   17.100 15.57
4  3B     4A 2014-09-01 13:00:00     1.46    6.533  8.00
5  2A     10 2014-09-01 14:00:00     8.31   26.283 26.29
6  5B     4C 2014-09-01 12:00:00     1.04    8.583  8.00
```

FIGURE 1 Data Snapshot

After removing discrepant data (such as trips with zero fares), we moved on to preliminary exploratory analysis, fare and distance distributions, finally diving head-on into the deep end of the statistical pool.

3 | THE GAUSSIAN PROCESS

The Gaussian process is a stochastic process for which any collection of random variables follows a multivariate normal distribution. Gaussian processes are particular useful for modeling continuous functions [1]. To indicate that a function f follows a Gaussian process, we write $f \sim \mathcal{GP}(m, k)$. Here m represents a mean function and k represents a covariance function. From a Bayesian perspective, a Gaussian process can be seen as a distribution over a class of continuous functions which makes it a natural choice for a prior on an unknown function underlying some data generating process.

Let $x(t)$ represent the observed value at some point t for some data generating process. More specifically, we can let $x(t) = f(t) + \epsilon(t)$ where $\epsilon(t)$ are Gaussian errors with mean zero and variance σ^2. Thus, observations $x(t)$ are realizations from the true underlying, data generating process $f(t)$ plus some random noise. If we say f follows a Gaussian process with mean function m and covariance function k then this is equivalent to saying that for any finite collection of n points from f, $(f(t_1), f(t_2), \ldots, f(t_n))$,

$$(f(t_1), f(t_2), \ldots, f(t_n)) \sim \mathcal{N}((m(t_1), m(t_2), \ldots, m(t_n)), K(t_1, t_2, \ldots, t_n))$$

. This can be more compactly written as

$$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{m}, K).$$

which means that $\boldsymbol{x} = \boldsymbol{f} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$

The usefulness of this formulation comes from the properties of normal distributions. If $\boldsymbol{x}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{m}, K)$, then

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{f} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2 \boldsymbol{I} + K & K \\ K & K \end{bmatrix} \right)$$

which means that the marginal distribution of $\boldsymbol{x}$ is

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{I} + K)$$

and the posterior distribution of $\boldsymbol{f}$ given the collection of observations $\boldsymbol{x}$ is

$$\boldsymbol{f}|\boldsymbol{x} \sim \mathcal{N}(K(\sigma^2 \boldsymbol{I} + K)^{-1}(\boldsymbol{x} - \boldsymbol{m}), K - K(\sigma^2 \boldsymbol{I} + K)^{-1} K).$$

Similar properties can be exploited to find the posterior predictive distribution of $\tilde{\boldsymbol{f}}$ for a new collection of values $\tilde{\boldsymbol{x}}$.

In most cases, since little prior knowledge is available for $\boldsymbol{m}$, it is assumed that $\boldsymbol{m} = \mathbf{0}$. Thus, the marginal distribution of $\boldsymbol{x}$ is

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I} + K)$$

and the posterior distribution for $\boldsymbol{f}$ is

$$\boldsymbol{f}|\boldsymbol{x} \sim \mathcal{N}(K(\sigma^2 \boldsymbol{I} + K)^{-1} \boldsymbol{x}, K - K(\sigma^2 \boldsymbol{I} + K)^{-1} K).$$

3.1 | Covariance Functions

When $\boldsymbol{m} = \mathbf{0}$, the possible class of functions for $\boldsymbol{f}$ is completely determined by the covariance function k. There are several popular choices for covariance functions and a host of other covariance functions can be formed by taking products or sums of these more standard covariance functions. In this paper, for reasons that will become obvious later on, we will only discuss two covariance functions: the squared exponential covariance function as well as the periodic covariance function.

The squared exponential covariance function is given by

$$k(t,t') = \tau^2 \exp\left(- \frac{|t-t'|^2}{2l^2} \right).$$

The parameter τ^2 is the signal variance and the parameter l^2 is the length-scale parameter. The τ^2 controls the height of the function and l^2 controls how wiggly the function is. More intuitively, l can be thought of as how close two points have to be to significantly influence each other. Large values of l produce very smooth functions and small values of l produce erratic, but still smooth, functions. This covariance function is very flexible and can be used to represent a wide variety of functions. In fact, by choosing l to be sufficiently small, almost any function can be represented accurately, but this might result in overfitting.

The periodic covariance function is given by

$$k(t,t') = \tau^2 \exp\left(- \frac{2\sin^2(\pi|t-t'|/p)}{l^2} \right).$$

The parameter τ^2 has the same function and interpretation as before, as does the l^2 parameter. However, there is an additional parameter p that represents how often the function repeats. This function is still very flexible, although not as flexible as the squared exponential covariance function since it should only be used for processes that are periodic. Still, choosing a sufficiently small l can lead to overfitting.

4 | A GAUSSIAN PROCESS FOR DAILY UBER TRIPS

A Gaussian process is used to help model the number of daily Uber trips over a period of a year, from September 2014 to August 2015. Let $Y(t)$ represent the number of trips made on day t for $t = 1, 2, \ldots, N$ where $N = 365$. As is common practice when using Gaussian processes, the $Y(t)$ are centered and scaled to fit the model. The t are also centered. We choose to model $Y(t)$ as follows:

$$Y(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + f_1(t) + f_2(t) + \epsilon(t) \text{ where } \epsilon(t) \sim \mathcal{N}(0, \sigma^2).$$

Here β_1 and β_2 are coefficients corresponding to linear and quadratic trends respectively. These are included in the model to account for the fact that the number of Uber trips might be growing over the year as the company grows, but this doesn't reflect a long term, yearly trend. This long term yearly trend is modeled with $f_1(t)$. A periodic, weekly trend is modeled with $f_2(t)$.

If we let $X = [\mathbf{1}, \boldsymbol{t}, \boldsymbol{t}^2]$, then the model can be written more compactly as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{f}_1 + \boldsymbol{f}_2 + \epsilon \text{ where } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I).$$

The functions $\boldsymbol{f}_1, \boldsymbol{f}_2$ are given Gaussian process priors. Thus, any finite collection of values of these functions are modeled as normal distributions, both with mean vector $\mathbf{0}$, whose covariance matrices are computed using squared exponential and periodic covariance functions respectively. More specifically, we have

$$\boldsymbol{f}_1 \sim \mathcal{N}(\boldsymbol{0}, \tau_1^2 K_1), \; \boldsymbol{f}_2 \sim \mathcal{N}(\boldsymbol{0}, \tau_2^2 K_2)$$

where

$$K_1(t, t') = \exp\left(- \frac{|t - t'|^2}{2 l_1^2} \right), \; K_2(t, t') = \exp\left(- \frac{2 \sin^2(\pi |t - t'|/p)}{l_2^2} \right).$$

Currently, we have chosen to fix the parameters l_1^2, l_2^2, and p to reasonable values. Since we expect a weekly, periodic trend, we let $p = 7$. Both l_1^2 and l_2^2 are also fixed at 7^2 meaning that observations more than 7 days apart will have little influence on each other.

For priors on the model parameters, we chose:

$$\begin{aligned}
&\pi(\boldsymbol{\beta}) \propto 1 \\
&\sigma^2, \tau_1^2 \sim \text{InverseGamma}(3, 5) \\
&\tau^2 \sim \text{Gamma}(1.5, 0.2)
\end{aligned}$$

Values for hyperparameters for $\sigma^2, \tau_1^2, \tau_2^2$ were chosen to be weakly informative, with large prior variances and most of the mass concentrated on values above 0 and below 100. The prior for $\boldsymbol{\beta}$ is also weakly informative. Normally, this prior for β_0 would be a poor choice, but the data have been standardized and the days have been centered, we don't expect β_0 to be large, nor do we expect β_1 and β_2 to be large.

4.1 | Gibbs Sampling Scheme for Daily Uber Trips Model

Posterior samples for $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ $\boldsymbol{\beta}$, σ^2, τ_1^2, and τ_2^2 are obtained using a Gibbs sampling scheme with an MH step for τ_2^2. Since $\boldsymbol{Y} | \boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(X\boldsymbol{\beta} + \boldsymbol{f}_1 + \boldsymbol{f}_2, \sigma^2 \boldsymbol{I})$, a single iteration in the sampling scheme is as follows:

1. Sample $\boldsymbol{\beta}| \cdot$ from $\mathcal{N}((X'X)^{-1} X'\boldsymbol{Y}, \sigma^2 (X'X)^{-1})$
2. Sample $\sigma^2 | \cdot$ from InverseGamma$(3 + \frac{N}{2}, 5 + 0.5(\boldsymbol{Y} - X\boldsymbol{\beta} - \boldsymbol{f}_1 - \boldsymbol{f}_2)'(\boldsymbol{Y} - X\boldsymbol{\beta} - \boldsymbol{f}_1 - \boldsymbol{f}_2))$
3. Sample $\tau_1^2 | \cdot$ from InverseGamma$(3 + \frac{N}{2}, 5 + 0.5\boldsymbol{f}' K_1^{-1} \boldsymbol{f})$
4. Propose $(\tau_2^2)^*$ from $\mathcal{N}(\tau_2^2, 1)$ and accept proposal with probability $\min\left\{1, \frac{f(\boldsymbol{f}_2 | (\tau_2^2)^*) \pi_{Gamma}((\tau_2^2)^*, 1.5, 0.2)}{f(\boldsymbol{f}_2 | \tau_2^2) \pi_{Gamma}(\tau_2^2, 1.5, 0.2)}\right\}$
5. Sample $\boldsymbol{f}_1$ from $\mathcal{N}((\tau_1^2 K_1)(\sigma^2 \boldsymbol{I} + (\tau_1^2 K_1) + (\tau_2^2 K_2))^{-1} (\boldsymbol{Y} - X\boldsymbol{\beta}), (\tau_1^2 K_1) - (\tau_1^2 K_1)(\sigma^2 \boldsymbol{I} + (\tau_1^2 K_1) + (\tau_2^2 K_2))^{-1} (\tau_1^2 K_1))$
6. Sample $\boldsymbol{f}_2$ from $\mathcal{N}((\tau_2^2 K_2)(\sigma^2 \boldsymbol{I} + (\tau_1^2 K_1) + (\tau_2^2 K_2))^{-1} (\boldsymbol{Y} - X\boldsymbol{\beta}), (\tau_2^2 K_2) - (\tau_2^2 K_2)(\sigma^2 \boldsymbol{I} + (\tau_1^2 K_1) + (\tau_2^2 K_2))^{-1} (\tau_2^2 K_2))$

This process was repeated 10,000 times with three chains run from different initial values. The first 1,000 draws were discarded.

4.2 | Results and Model Checking

To assess convergence and mixing for this model, we use several tools including the Gelman-Rubin diagnostic [2], effective sample size, and trace plots. In Table 1, we see that the Gelman-Rubin diagnostics are close to 1 for the parameters $\boldsymbol{\beta}, \sigma^2, \tau_1^2, \tau_2^2$ which supports our conclusion that we have converged to the posterior distribution. We also see the effective sample sizes for these parameters. Except for τ_2^2, the sample sizes are relatively large which suggests that the draws weren't too heavily correlated.

	GR Estimate	GR Upper CI	ESS
β_0	1.0001	1.0006	26641.7164
β_1	1.0001	1.0005	23901.4395
β_2	1.0002	1.0007	36927.7226
σ^2	0.9999	0.9999	25271.5523
τ_1^2	1.0004	1.0018	11983.7929
τ_2^2	1.0172	1.0278	98.0828

TABLE 1 Gelman-Rubin diagnostics and effective samples sizes for model of daily Uber trips

In Figure 2 we see the trace plots for the different parameters. These plots provide further evidence that our chains have converged to the posterior distribution and that they mixed relatively well.

To assess model adequacy, we first we found the posterior mean for $X\boldsymbol{\beta} + \boldsymbol{f}_1 + \boldsymbol{f}_2$ and compared this to the observed values for $\boldsymbol{Y}$. Figure 3 shows how the posterior mean line fits the data. The black line shows the true observations and the red line shows the posterior mean line. We can see that the line follows the general pattern of the data, although there are some places in which we expected rides to increase, but they decreased. This usually occurred around holidays.

To investigate model adequacy further, we performed a posterior predictive check. For each draw from our joint posterior, we produced a new data set $\boldsymbol{Y}_{(i)}^{\text{pred}}$ as well as a draw from the posterior of $X\boldsymbol{\beta} + \boldsymbol{f}_1 + \boldsymbol{f}_2$, denoted here as $\boldsymbol{\theta}_{(i)}$. These samples were produced at the current values of t. For each $\boldsymbol{\theta}_{(i)}$, we computed both $MG(\boldsymbol{Y}, \boldsymbol{\theta}_{(i)}) = \max(\boldsymbol{Y} - \boldsymbol{\theta}_{(i)})$ and $MG(\boldsymbol{Y}_{(i)}^{\text{pred}}, \boldsymbol{\theta}_{(i)}) = \max(\boldsymbol{Y}_{(i)}^{\text{pred}} - \boldsymbol{\theta}_{(i)})$. A scatterplot of $MG(\boldsymbol{Y}, \boldsymbol{\theta}_{(i)})$ vs $MG(\boldsymbol{Y}_{(i)}^{\text{pred}}, \boldsymbol{\theta}_{(i)})$ can be found in Figure 4. An adequate model would see the red line going right through the center of the scatterplot. We can see that this in fact is the case for this model. This suggests that at least in terms of this measure, the model performs well.

4.3 | Estimated Trends

The purpose of fitting this model was to gain insight into the yearly and weekly trends in Uber trips. In Figure 5 we can see the different estimated components from the data. The first graph shows that demand in general is increasing over time. This is despite the fact that this period of time was one of financial distress for Uber. In fact, this trend shows one of the reasons why it was a time of such financial distress. The demand for drivers was rising and Uber wasn't making any money after paying the drivers for their services.

The second graph shows a yearly trend. This confirms that around certain holidays, demand for Uber drivers drops dramatically. It also shows when the highest demand occurred which was between Thanksgiving and Christmas and near the end of February and beginning of March.

In Figure 6 we get a better look at the periodic, weekly trend. As can be seen, the beginning of the week doesn't see high demand, but the weekends do. This suggests that Uber drivers will see the most demand for the services on the weekends.

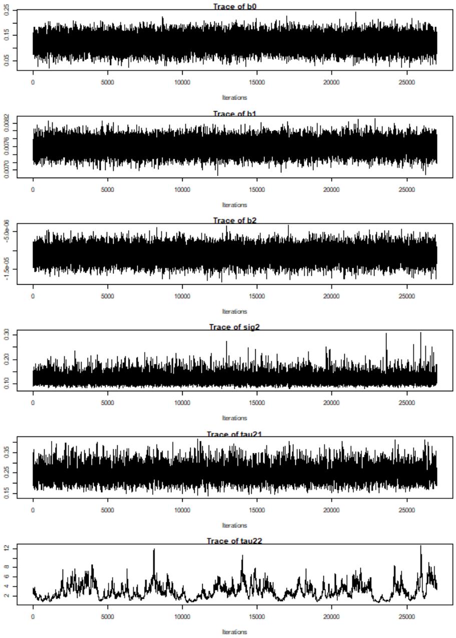


FIGURE 2 Traceplots of parameters for model of daily Uber trips

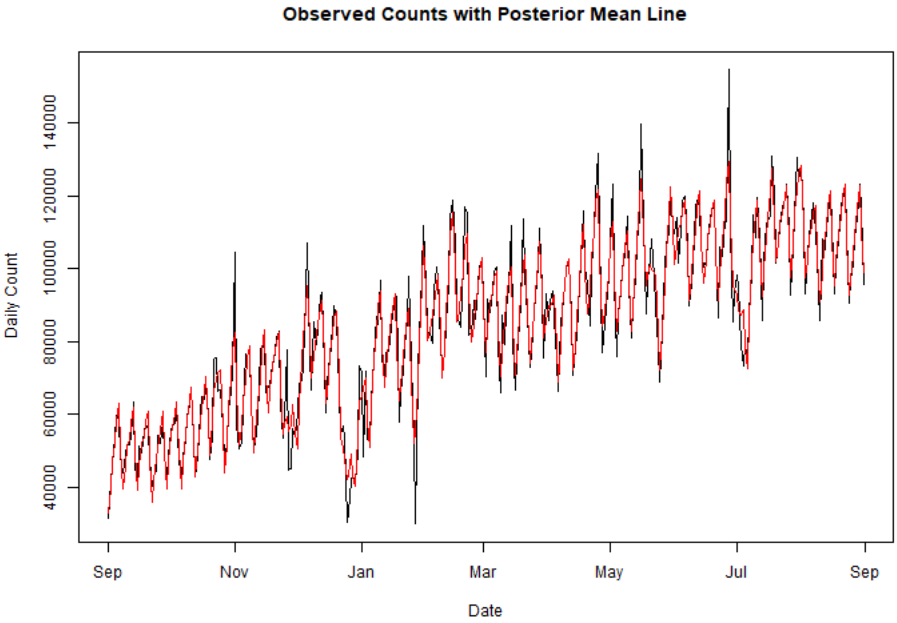


FIGURE 3 Predictions with observed values for model of daily Uber trips

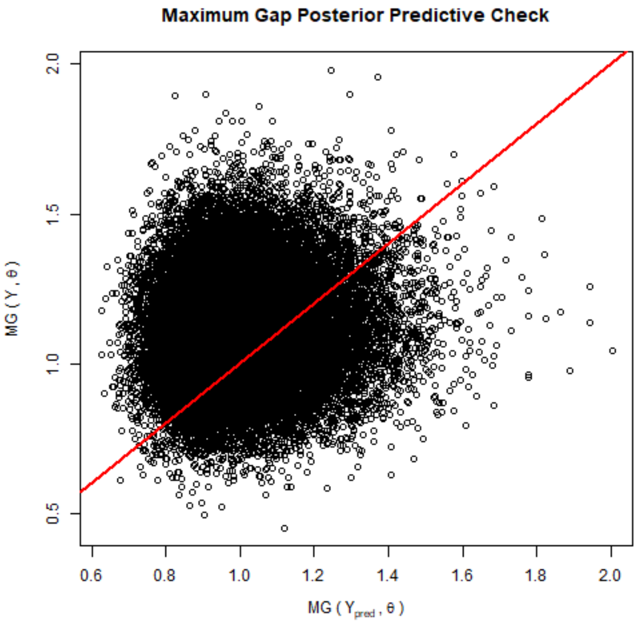


FIGURE 4 Posterior predictive check

5 | A SPATIO-TEMPORAL MODEL FOR HOURLY UBER TRIPS

Beyond knowing which times of year and week have the highest demand, it would be useful to know which hours of the day and which locations are the busiest. The model from the previous section could be used for individual locations in New York City, but it is convenient to be able to model relationships in space and time together instead of separately.

The Gaussian process model can be extended to model spatio-temporal phenomenon. Let $Y(s, t)$ represent the centered and scaled logarithm of the number of trips at location s and time t. There are fifteen locations and twenty-four times points at each location so then $S = 15$, $T = 24$ and $N = S \times T = 360$ is the total number of observation. We model $Y(s, t)$ as follows:

$$Y(s, t) = f(s, t) + \epsilon(s, t) \text{ where } \epsilon(s, t) \sim \mathcal{N}(0, \sigma^2).$$

Here $f(s, t)$ represents a spatio-temporal process. This model can again be written more compactly as

$$\boldsymbol{Y} = \boldsymbol{f} + \epsilon \text{ where } \epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I).$$

The spatio-temporal process $\boldsymbol{f}$ is modeled as a Gaussian process with mean function 0. The covariance function $\tau_s^2 \tau_t^2 k(s, s', t, t')$ is a function of both space and time. An easy thing to do is to treat the spatial and temporal components as separable and let $k(s, s', t, t') = k(s, s')k(t, t')$. For our model we let both $k(s, s')$ and $k(t, t')$ be squared

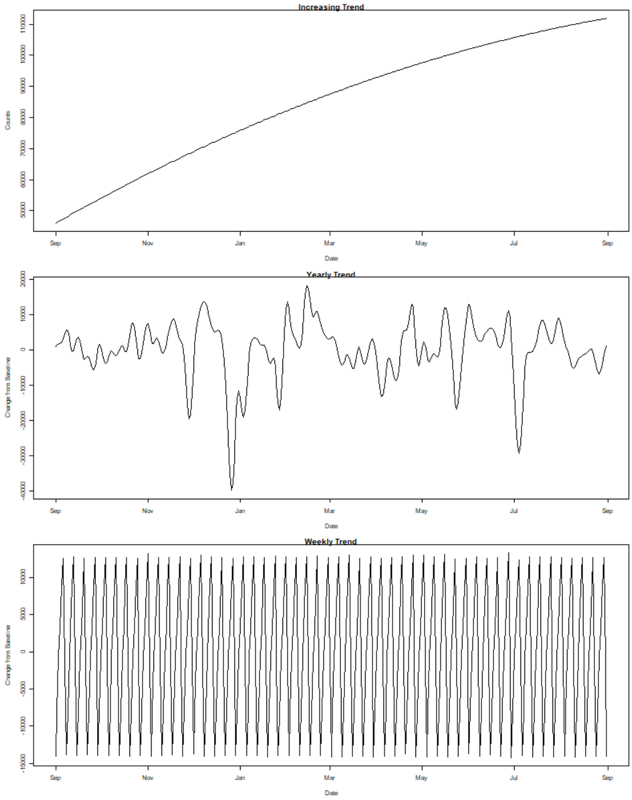


FIGURE 5 Increasing, yearly, and weekly trend components of daily Uber trips model

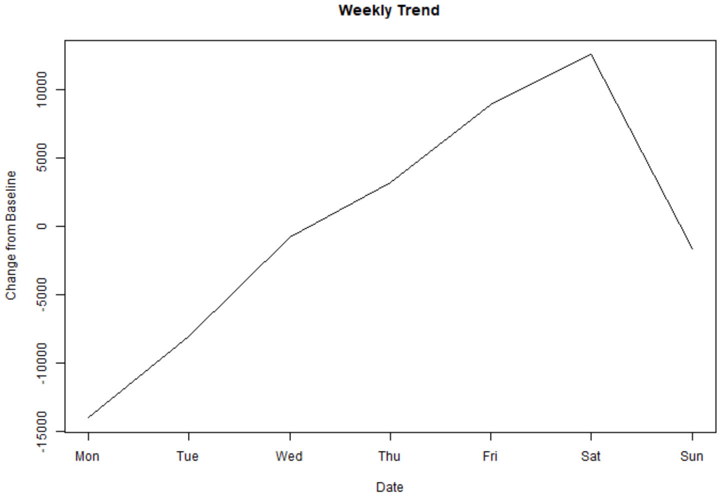


FIGURE 6 Weekly trend component of daily Uber trips model

exponential functions. Thus for any finite collection of points in space and time, we have

$$\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \tau_s^2 K_S \otimes \tau_t^2 K_T).$$

Once again, K_S and K_T are functions of length scale parameters which we have chosen to fix at $l_s^2 = 0.005$ and $l_t^2 = 12$. This is a small value for l_s which might make our spatial component too flexible, but it's important to remember that l_s represents how close locations need to be before they start to influence each other. Since location coordinates will be given in latitude and longitude and because locations are only found in New York City, Euclidean distances will be relatively small. If $l_s^2 = 0.005$, then l_s is about 0.071 which translates to about half the distance from Brooklyn to Queens.

As priors for model parameters we chose:

$$\sigma^2, \tau_s^2, \tau_t^2 \sim \text{InverseGamma}(3, 5)$$

Choosing these priors makes sampling from the posterior distribution particularly easy.

5.1 | Gibbs Sampling Scheme for Spatio-Temporal Model

Kronecker products have the following properties:

1. $(cK_1) \otimes K_2 = K_1 \otimes (cK_2) = c(K_1 \otimes K_2)$
2. $(K_1 \otimes K_2)^{-1} = K_1^{-1} \otimes K_2^{-1}$
3. $|K_1 \otimes K_2| = |K_1|^T |K_2|^S$

By using these properties, we can show that

$$|\tau_s^2 K_S \otimes \tau_t^2 K_T|^{-1/2} \exp(-0.5 \boldsymbol{f}' (\tau_s^2 K_S \otimes \tau_t^2 K_T)^{-1} \boldsymbol{f})$$

can be rewritten as

$$(\tau_s^2)^{-N/2} |K_S|^{-T/2} (\tau_t^2)^{-N/2} |K_T|^{-S/2} \exp(-0.5 \tau_s^{-2} \boldsymbol{f}' (K_1 \otimes \tau_t^{-2} K_2^{-1}) \boldsymbol{f})$$

or

$$(\tau_s^2)^{-N/2} |K_S|^{-T/2} (\tau_t^2)^{-N/2} |K_T|^{-S/2} \exp(-0.5 \tau_t^{-2} \boldsymbol{f}' (\tau_s^{-2} K_1 \otimes K_2^{-1}) \boldsymbol{f})$$

and it becomes obvious that the inverse gamma distribution is a conjugate prior for τ_s^2 and τ_t^2.

Therefore, the sampling scheme is as follows:

1. Sample $\sigma^2|\cdot$ from InverseGamma$(3 + \frac{N}{2}, 5 + 0.5(\boldsymbol{Y} - \boldsymbol{f})'(\boldsymbol{Y} - \boldsymbol{f})$
2. Sample $\tau_s^2|\cdot$ from InverseGamma$(3 + \frac{N}{2}, 5 + 0.5 \boldsymbol{f}' (K_1 \otimes \tau_t^{-2} K_2^{-1}) \boldsymbol{f})$
3. Sample $\tau_t^2|\cdot$ from InverseGamma$(3 + \frac{N}{2}, 5 + 0.5 \boldsymbol{f}' (\tau_s^{-2} K_1 \otimes K_2^{-1}) \boldsymbol{f})$

This was repeated on three different chains with 10,000 draws for each chain where the first 1,000 draws were

	GR Estimate	GR Upper CI	ESS
σ^2	1.0000	1.0001	16561.97
τ_s^2	1.0200	1.0571	2727.01
τ_t^2	1.0193	1.0517	1115.583

TABLE 2 Gelman-Rubin diagnostics and effective samples sizes for spatio-temporal model

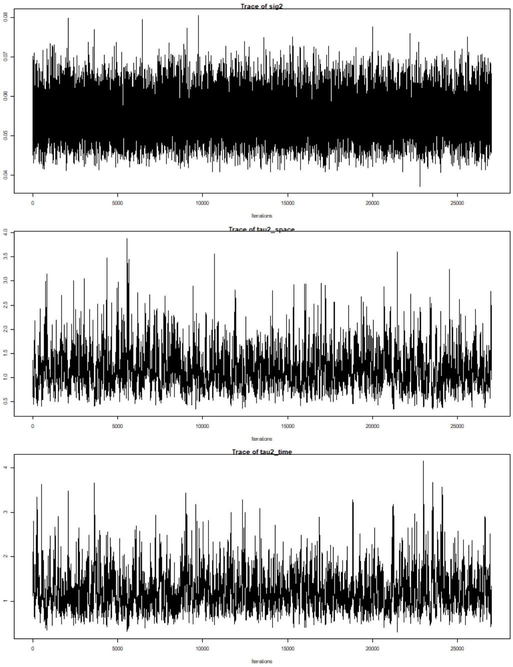


FIGURE 7 Traceplots of parameters for spatio-temporal model

discarded.

5.2 | Results and Model Checking

Convergence and mixing were again assessed by looking at trace plots as well as Gelman-Rubin diagnostics and effective sample sizes. Table 2 shows the Gelman-Rubin diagnostics as well as the effective sample sizes for σ^2, τ_s^2 and τ_t^2. We can see that these diagnostics suggest that the chains have converged and autocorrelation isn't too strong. By looking at the trace plots in 7, we see visually that the chains appear to mix well and have converged to the posterior distribution.

Model adequacy for this model was assessed by again comparing the posterior mean of $\boldsymbol{f}$ with the observed $\boldsymbol{Y}$ as well as by performing a posterior predictive check. From Figure 8, we can see that the model fits the data well. The observed lines in black are followed closely by the predicted lines in pink. In general we can see that fewer Uber trips occur in the morning and then are fairly constant during the day with a small bump in the early evening. This

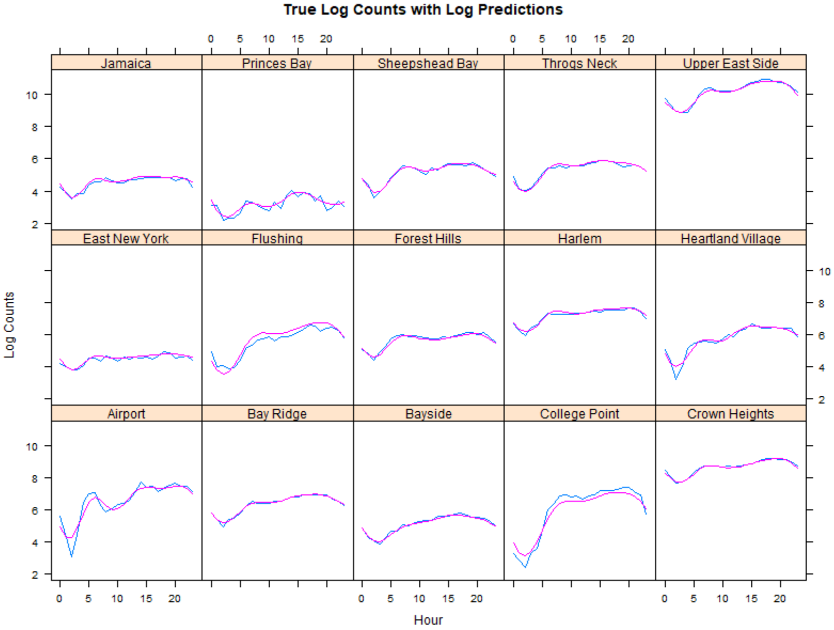


FIGURE 8 Predictions with observed values for spatio-temporal model

general trend is found almost everywhere although in some areas it is more pronounced. We can also confirm that areas closer to the center of New York City are busier.

To check how data generated from our model compares to the true observed data, we performed a posterior predictive check. For each draw from the posterior, we produced a new data set $\boldsymbol{Y}_{(i)}^{\text{pred}}$. We also obtained a sample of the latent spatio-temporal process $\boldsymbol{f}_{(i)}$. We then computed $Var(\boldsymbol{Y}, \boldsymbol{f}_{(i)}) = \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{Y}_j - \boldsymbol{f}_{(i)j})^2$ along with $Var(\boldsymbol{Y}_{(i)}^{\text{pred}}, \boldsymbol{f}_{(i)}) = \frac{1}{N}\sum_{j=1}^{N}((\boldsymbol{Y}_{(i)j}^{\text{pred}} - \boldsymbol{f}_{(i)j})^2$ for each of these draws. We can see the scatterplot of these values in 9. Again, the model seems to be inadequate in this regard. It appears that there is more variability in the generated data than there is in the true data.

6 | UBER TRAFFIC AS AN DIRECTED EDGE WEIGHTED GRAPH

A complete digraph is a directed graph in which every pair of distinct vertices is connected by a pair of unique edges. Uber traffic in NYC comprises 29 unique sources(destinations), each of which are connected to one another via an edge, whose weight denotes the (median) fare charged by the Uber driver while driving from source to destination. The number of edges in the digraph is $n \cdot (n-1) = 29 \cdot 28 = 812$. The complete graph K_{29} displays all possible Uber trips in NYC. While K_{29} is an extremely compact visual representation of the thirty million trips, we must note that each edge in the graph corresponds to *all* the fares obtained by the driver while plying the edge. In particular, every edge corresponds to an entire fare distribution with hundreds of fares! The edge fare distribution may be approximated by a Gaussian, as we shall do so subsequently. However, this is merely a convenience to enable Bayesian analysis. Each edge fare distribution in actuality is a multimodal collection of dollar amounts that vary from a minimum $8 to over $75, depending on the route plied.

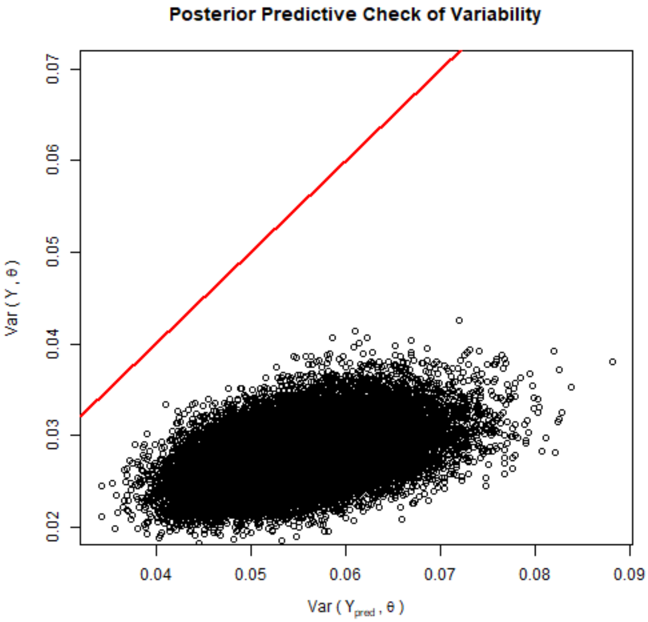


FIGURE 9 Posterior predictive check

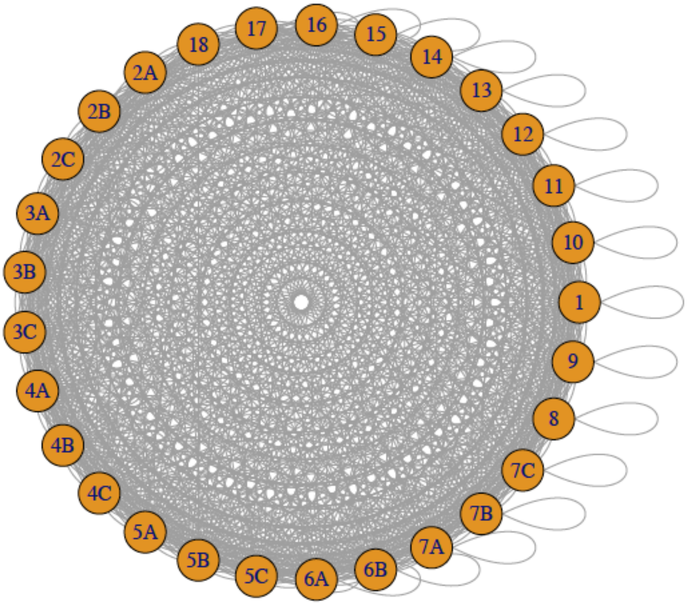
7 | THE OPTIMAL DAILY ROUTE

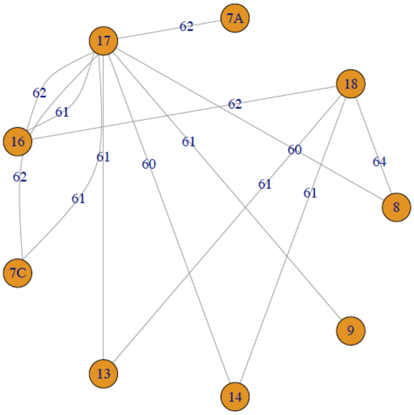
We seek an optimal route that makes the most money for the Uber driver. We impose several pragmatic constraints upon this open-ended problem to make it tractable. The typical full-time Uber driver works an eight hour shift during the day, averaging a trip every hour. Hence we seek a path with eight vertices. Furthermore, a closed loop enables the Uber driver to return to his starting vertex. This is highly preferable as the Uber driver may start at any one of the eight vertices along the loop in the morning, complete the loop and return to the starting spot at the end of the day with all trips paid for by the passenger; in the absence of a loop, the last leg will have to be paid for by the driver out of his pocket!

Finding a closed path of length eight such that the path fetches the highest amount of median fare; equivalently, finding a closed loop of fixed length where the edge weights sum to a maximum over all other loops of equal length, is an NP-hard problem. Given a graph, the graph library *igraph* enumerates all simple paths from a source to a target vertex. However, the library cautions "there are exponentially many paths between two vertices of a graph, and you may run out of memory when using this function". Indeed, the number of paths between two vertices in a complete graph of moderate size is in the tens of millions! Further, since two vertices can be chosen in $\binom{n}{2}$ ways, the number of simple paths exceed a billion! Morover, the length of each of these billion paths is not known a-priori! Devising a strategy to find a loop of length eight turned out to be much harder than we anticipated.

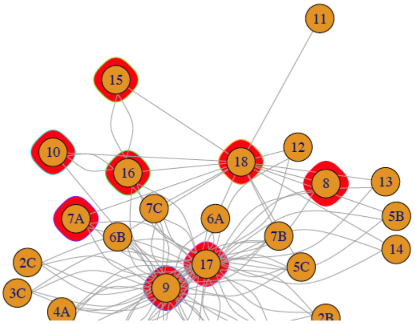
An iterative strategy was finally hit upon. We start with an extremely sparse graph, obtained by retaining only those edges that fetch a median $60 or more.

Uber as a graph, with 29 vertices, 812 edges.

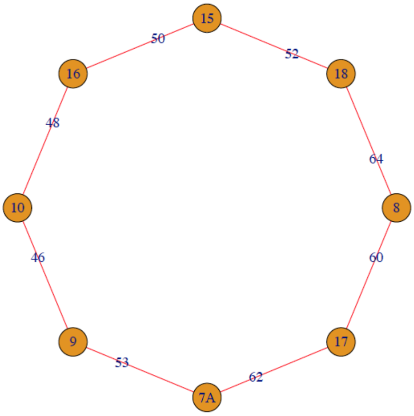




If the desired path is not found therein, we relax the fare restriction, examining graphs with median fares of $59, and so on. This numerically intensive computation finally yields a closed loop of length eight when the dollar amount is $46. The nodes participating in this loop are displayed in red below.

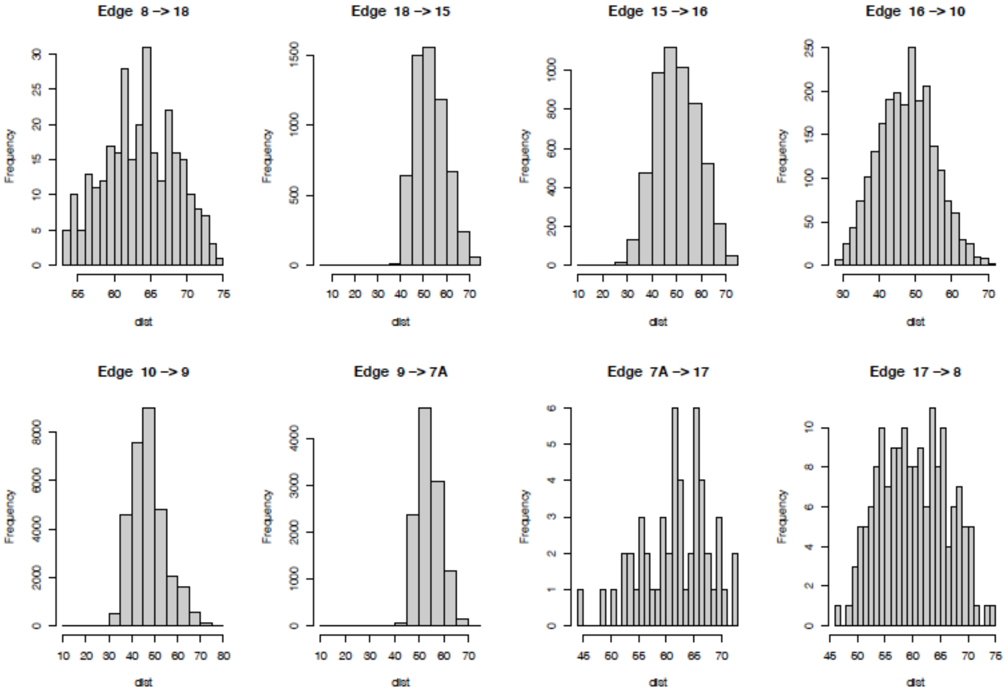


It must be noted that this lucrative 8-gon is just one of several; it isn't the most lucrative 8-gon - to find such an 8-gon would require enumerating all the simple paths of the original K_{29}, which is beyond the capabilities of a regular consumer laptop. The desired closed loop is shown below. The edge weights denote the median fare obtained when the Uber driver plies the route along the edge. Summing the edge weights gives an estimate of the median fare obtained by plying the closed circuit.



8 | OPTIMAL CIRCUIT FARES:BDA FOR NORMAL DATA, UNKNOWN μ

Perhaps the simplest approach, as suggested by examining the histograms of the edge fares below, is to model each edge distribution as a Gaussian with a mean fare of about \$50. (We assume the standard deviation of these edges to be known and equal \$7, which is approximately the sample standard deviation for these edges). The circuit is then a sum of 8 independent Gaussians.



The statistical model: We attribute a Gaussian prior to the Optimal Circuit fare.

Conjugate prior $\mu \sim N(\mu_0, \tau_0^2) = N(8 * 50, 8 * 7^2)$

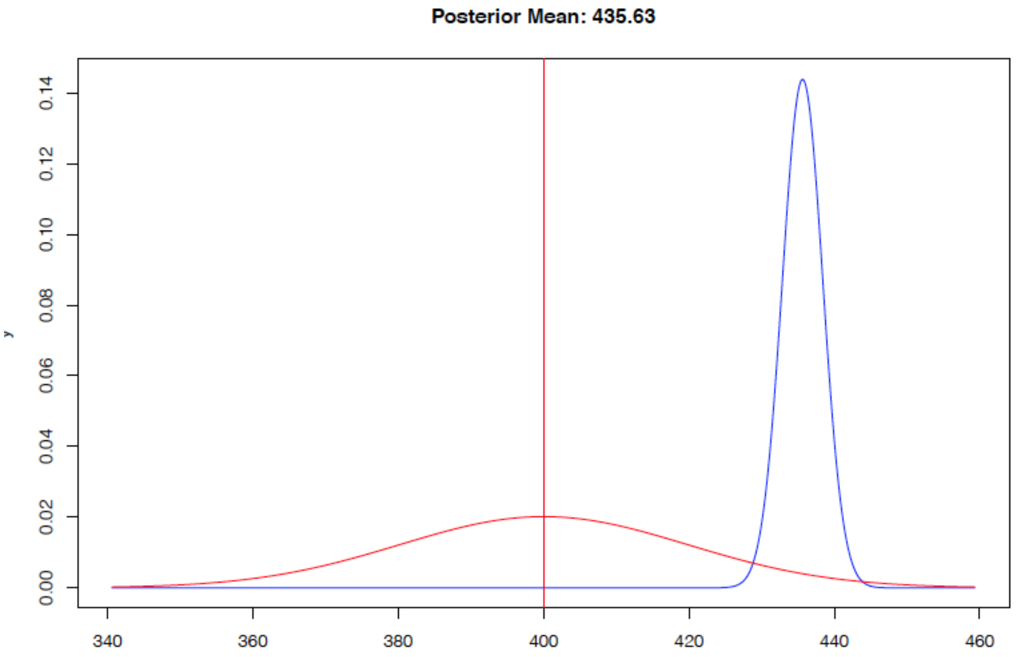
We now drive 50 times around this circuit. Everytime we drive around the circuit, we obtain an edge fare from each of the eight corresponding edge distributions. Eight such fares sum to one circuit fare. Armed with 50 circuit fares, we may compute the likelihood:

$$L(\mu|y_1, ...y_{50}) = L(\mu|\bar{y}) = exp\frac{-50(\mu-\bar{y})^2}{2\sigma^2}$$

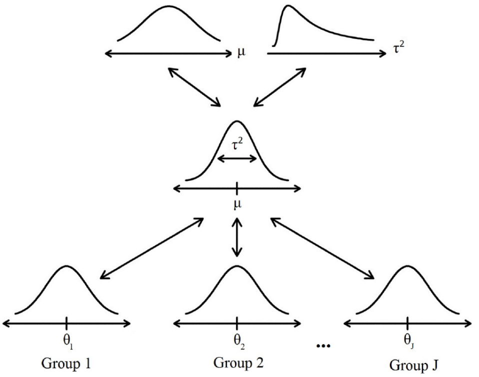
The posterior distribution of the mean fare of the Optimal Circuit, conditioned on the data, is given by:

$$\mu|\bar{y} \sim N(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2}\mu_0 + \frac{n/\sigma_0^2}{1/\tau_0^2 + n/\sigma^2}\bar{y}, (1/\tau_0^2 + n/\sigma^2)^{-1})$$

The prior in red and the posterior in blue are displayed below. Conclusion: Driving around the optimal route nets the Uber driver $435 a day. This amount is much higher than routes of similar length (same number of vertices).



9 | OPTIMAL CIRCUIT FARES: BDA WITH NORMAL HIERARCHICAL MODEL



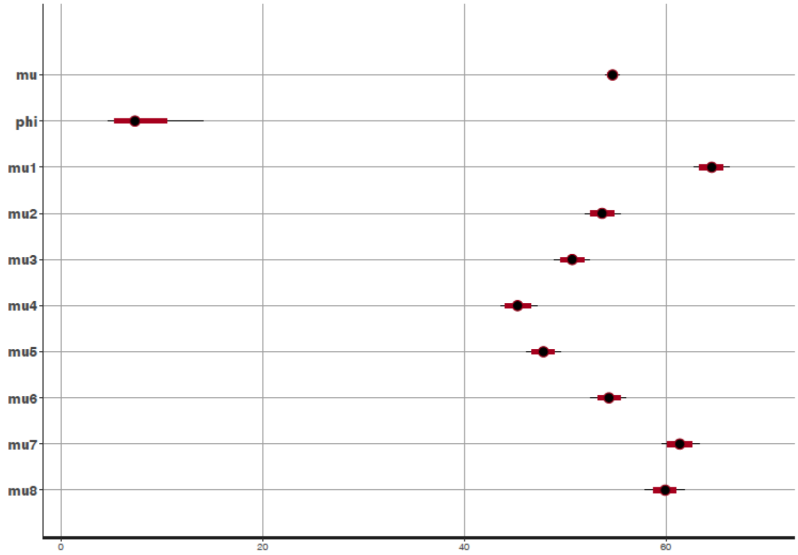
Using a plug-in estimate of $50 as the mean fare per edge ignores uncertainity about the parameter, while also being inaccurate as each of these edges comprise a different route in the city, with differing mean fares. A Normal hierarchical model allows for varying mean fares across the eight edges. The key feature is that the circuit prior mean is not presumed to be a known quantity of $400. Instead, the circuit prior is unknown and has a its own distribution, a hyper-prior. The model is given by:

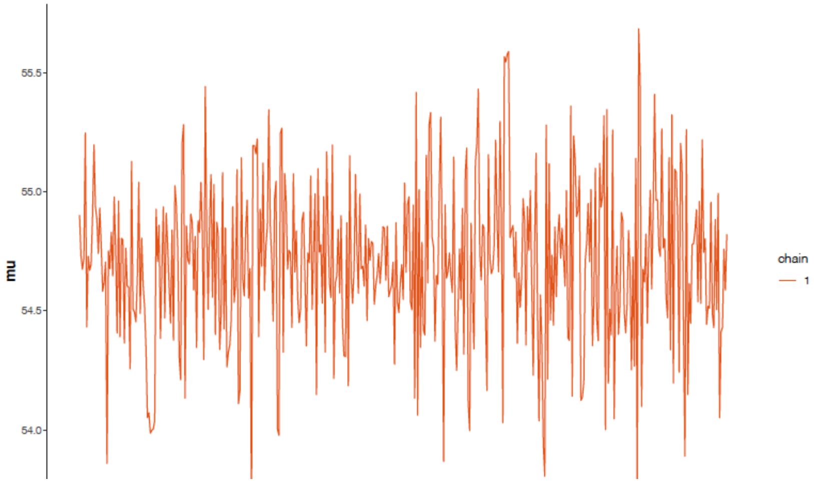
1. $y_{ij}|\theta, \mu, \tau^2 \sim N(\theta_j, \sigma^2), i = 1..n_j, j = 1..8$

2. $\theta_j|\mu, \tau^2 \sim N(\mu, \tau^2)$

3. $p(\mu, \tau^2) = p(\mu|\tau^2)p(\tau^2)$

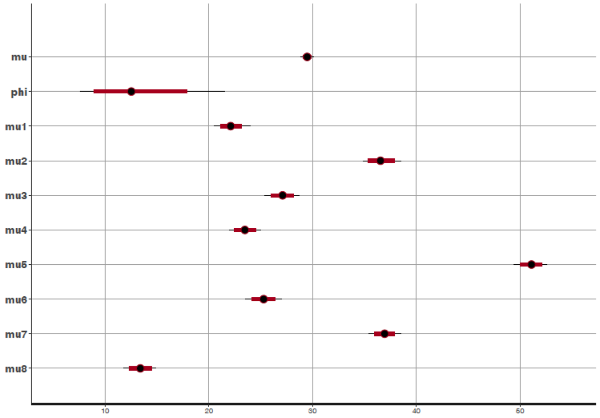
The hierarchical model is defined via Stan and computed analytically using MCMC samplers. Results are shown below. The posterior mean of the optimal circuit differs from the previous estimate by less than five dollars (<1%).

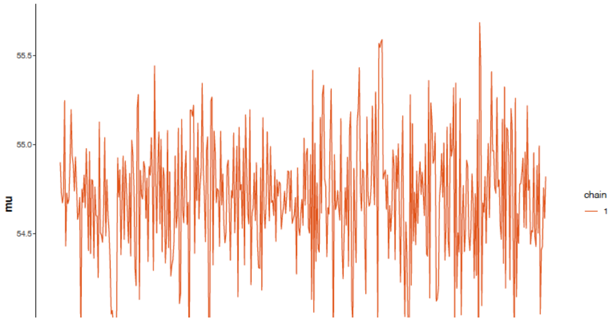




10 | RANDOM CIRCUIT FARES: NORMAL HIERARCHICAL MODEL

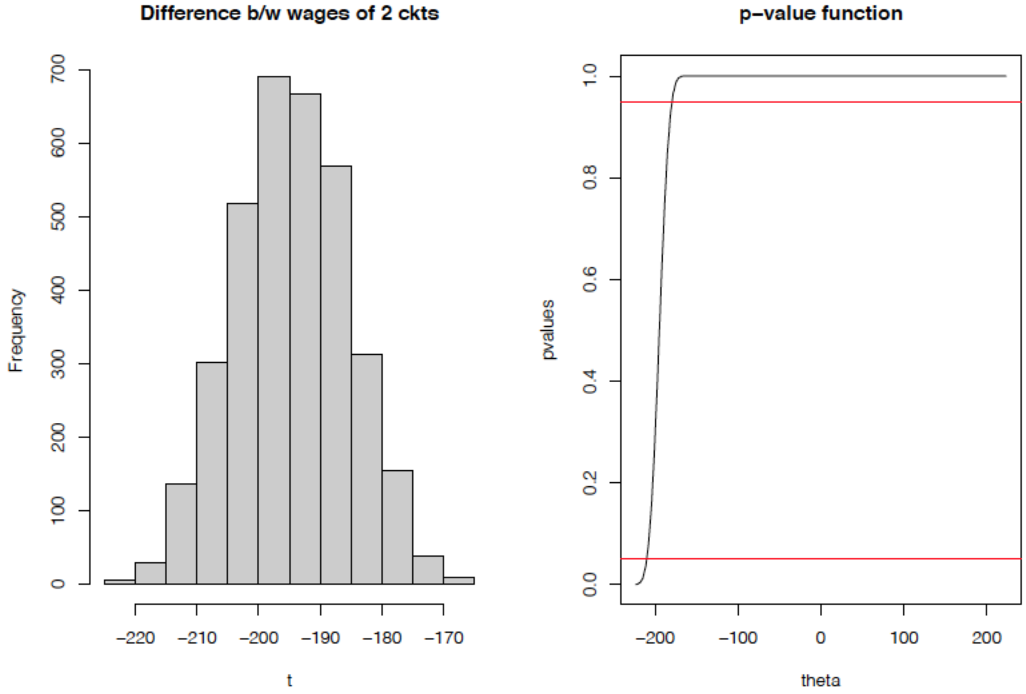
A rookie cabbie would execute a random walk around the Uber graph, essentially picking up the first passenger he gets, dropping him to his destination, picking up the next passenger at that point and so forth. While Uber perhaps hopes that its cabbies behave in this random fashion to ensure uniform service, it is not in the interest of the Uber driver to do so. The median fare of a random (sub-optimal) daily route is estimated via the Normal Hierarchical Model. The take-home estimate is a dismal $235! The foolish cabbie throws away a full $200 every single day while plying this sub-optimal route! A little statistical analysis goes a LONG way!





11 | TELLING APART UBER TRIPS: THE CONFIDENCE DISTRIBUTION

Professor Don Rubin characterizes the Fisher Randomization Test as a Stochastic Proof by Contradiction. FRT is a rather computationally intensive tool, making a comeback as a Distribution-Free alternative to comparing two messy multi-modal distributions. A Boolean assignment vector over two weeks assigns a driver to drive either the Optimal circuit A or another (sub-optimal) circuit B. The difference of biweekly averages provides a simple test statistic, visualized via a p-value function. Thus, the fare distributions of two complex Uber circuits may be compared without making grossly simplifying assumptions about their underlying structure (such as fares belonging to a unimodal Gaussian). The authors thank Dr. Tirthankar Dasgupta for adding this deceptively powerful tool into their statistical arsenal.



12 | CONCLUSION

As ridesharing services like Uber become more popular in an ever growing gig economy where individuals are looking for a steady second income, knowing when and where to go to look for customers is vital. We've shown that the number of daily Uber trips can be modeled with a Gaussian process to reveal times of the year and week that are the busiest for Uber drivers in New York City. We found that not many individuals are using Uber during holidays, but a busy time is between Thanksgiving and Christmas. Additionally, weekends tend to be the busiest where trips reach their peak on Saturday. A Gaussian process was also used to model the number of hourly trips across time and space simultaneously. This model confirmed that areas in the center of New York City are the busiest and that peak hours are in the morning and early evening.

Another concern for Uber drivers is whether or not it matters to find an 'optimal' circuit or if they can drive around randomly looking for rides. Our analysis showed that for what we considered to be the optimal circuit, an Uber driver could expect to make around $440 dollars in a single day, which was significantly higher than what one would make when following a non-optimal, random circuit which would only allow them to make around $200 dollars. While it's not in Uber's interest to advertise the existence of these optimal routes, its definitely in the Uber driver's self-interest to seek out an optimal route in order to maximize his personal income.

While these models are immensely useful, there is room for improvement. Currently, the length-scale parameters in the Gaussian processes are fixed and are not being estimated. To some extent, these parameters control whether or not underfitting or overfitting occurs and so it makes sense to try and find their optimal values. A common approach is to estimate them from the marginal posterior obtained from integrating out the latent Gaussian process(es). Another component also needs to be included in the model of daily Uber trips to account for unusual patterns during holidays. Finally, a Bayesian method to find Optimal routes while incorporating temporal conditions is most sought after. The independence assumption for edge fares is also not robust. Covariances among edge fares can and do affect the optimal circuit fare. Nevertheless, these results provide a powerful first step to maximizing the Uber driver's daily take-home fare.

References

[1] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian data analysis*. CRC press.

[2] Gelman, A., Rubin, D. B. et al. (1992) Inference from iterative simulation using multiple sequences. *Statistical science*, **7**, 457–472.