Predictive power:
Kriging and Machine Learning

Antonio Jurlina
SIE 512

Predictive power:
Kriging and Machine Learning

Antonio Jurlina
SIE 512
Fall 2020
Dr. Kate Beard-Tisdale

I. Objective

The purpose of this project is to assess the predictive power of kriging (along with a generalized additive model) relative to that of a simple machine learning method like Random Forests, with ordinary least squares serving as the simplest (and most biased) approach. Furthermore, this project is specifically interested in the effect geographic location has on expected earnings for cab drivers in Manhattan. ***If drivers pick passengers up in downtown Manhattan, they will earn a higher wage on average, everything else equal, implying that geographic latitude is the most relevant factor when determining where to start.*** This is not a project that seeks to determine which specific variable contributes the most (on average with everything else held equal) to the final fares earned by cab drivers. Rather, it seeks to discover relevant relationship between latitude, longitude and fares collected. All other (known) effects are going to be accounted for and stripped away.

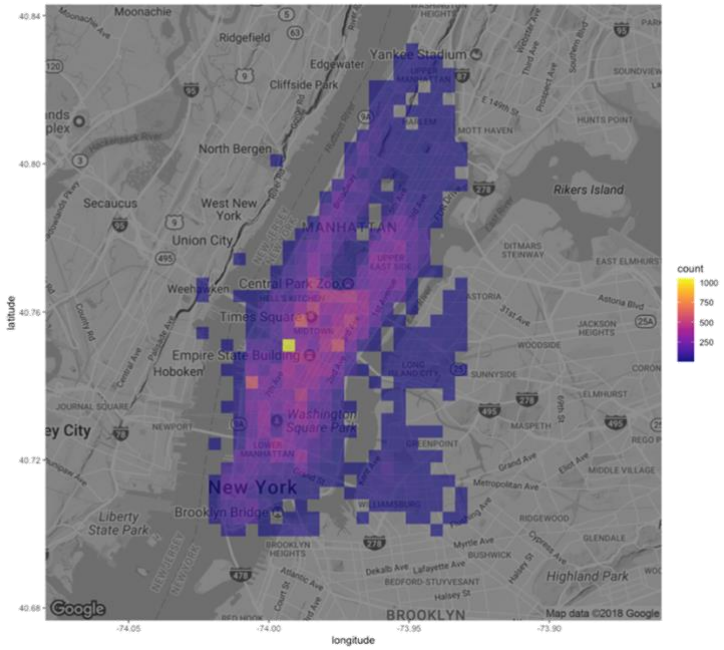

Figure 1: Spatial distribution of data

	log(total amount)	longitude	latitude	passenger count	trip distance	wday	hour	month	geometry	
1	3.23	-74	40.7	1	5.8	Wed	0	May	(-73.98457	40.71474)
2	1.7	-74	40.7	1	0.75	Thu	12	Aug	(-73.99466	40.74018)
3	2.2	-74	40.8	1	0.6	Thu	9	Feb	(-73.96581	40.75857)
4	2.01	-74	40.8	2	0.9	Sat	20	Feb	(-73.9655	40.79076)
5	2.56	-74	40.7	6	2.8	Sun	2	Oct	(-73.9968	40.74702)
6	2.31	-74	40.8	1	1.83	Mon	1	Feb	(-73.9706	40.76213)

Table 1: First 6 (out of 50,000) rows of the data

II. Data

The data set used for this project comes from the "FOILing NYC's Taxi Trip Data" project (Whong, 2014). Chris Whong published 20 gigabytes of NYC Taxi data which includes fares, tips, pickup and drop-off latitudes and longitudes, medallion IDs, passenger numbers, trip durations, and more (Table 1). The sheer size of the data set (more than a 160 million observations over 2013 and 2014) meant that it would be computationally too expensive trying to process it locally. Therefore, a subset of 50,000 observations was randomly sampled, with each month of each year providing a roughly equal share of data points. This hopefully preserved any temporal effects that the data might have. Relevant visualization statistics are included in Figures 1 and 2. There was no missing data and trips with zero total fares earned were excluded (as they represent extremely unlikely and illegal scenarios usually).

	log(total earnings)	passenger count	trip distance
Min	1.099	1	0
1st Quartile	2.079	1	1
Median	2.398	1	1.735
Mean	2.447	1.717	2.545
3rd Quartile	2.755	2	2.978
Max	5.08	6	53

Table 2: Summary statistics of non-categorical variables

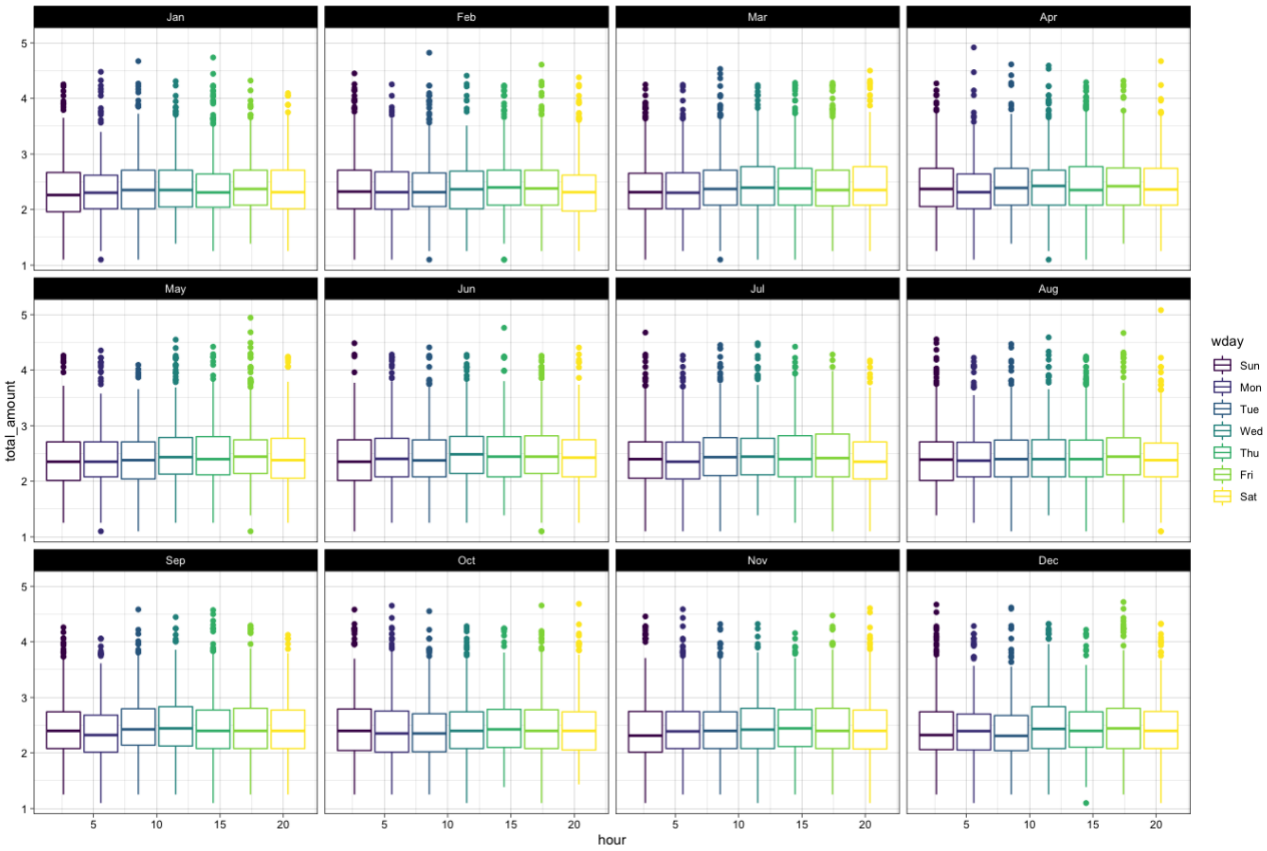

Figure 2: Temporal breakdown of the dependent variable

III. Methods

A simple OLS approach (Equation 1), in which the variable of interest is the log of total earnings (sum of the fare, tip, surcharge and tax), is the baseline approach of this project. Independent variables are pickup latitude, pickup longitude, trip distance, and number of passengers. Next, a GAM (Equation 2) is estimated over the same basic formula, but with smoothing applied to latitude, longitude and trip distance (with an assumption that the exact shape of the effect these variables have on final earnings is unknown). The use of the GAM (to be followed with kriging of residuals) comes from a very similar idea Gámez et al. (2000) had in their work on estimating housing prices in Albacete by incorporating the neighborhood effects

through spatial autocorrelation estimation and subsequent kriging of GAM residuals. In Manhattan, the assumption is that downtown pickups result in higher earnings. This indicates a certain weight neighboring points will have on each other as an isotropic spatial autocorrelation process. Therefore, the residuals of the GAM are expected to be biased, and kriging will be performed to account for and fix this. Kriging itself has a random component to it, as the underlying Gaussian process is used to interpolate predictions (Krige, 1951; Matheron, 1973; Matheron, 1963), which is what inspired the final aspect of this project, Random Forests.

$$\log(total\ earnings) = \alpha_0 + \alpha_1 * X_1 + \alpha_2 * X_2 + \alpha_3 * X_3 + \alpha_4 * X_4 + \alpha_5 * X_5 + \alpha_6 * X_6 + \alpha_7 * X_7 \quad (Equation\ 1)$$

$$\log(total\ earnings) = \alpha_0 + \alpha_1 * s(X_1) + \alpha_2 * s(X_2) + \alpha_3 * s(X_3) + \alpha_4 * X_4 + \alpha_5 * X_5 + \alpha_6 * X_6 + \alpha_7 * X_7 \quad (Equation\ 2)$$

Note: $X_1, X_2, X_3, X_4, X_5, X_6, X_7$ are pickup longitude, pickup latitude, trip distance, passenger count, day of the week, hour of the day, and month of the year, respectively; s() represents smoothing.

Much like kriging, Random Forest models (as the name implies) rely on an underlying random process and are used in estimating spatial models and making predictions (Breiman, 2001; Tin Kam Ho, 1998). In this project, a Random Forest model is used to estimate Equation 1, by growing 500 random decision trees across the data space and averaging them out into a prediction set. For the purposes of training the three models and subsequently testing them, the data is split into two subsets: first one contains a random subset of 75% of the data for training the models, while the second contains the remaining 25% for testing the predictions and estimating root mean square errors (Figure 3). These two datasets are used for the OLS, the GAM and kriging combination (with a spherical effect variogram model of the spatial dependence), and the Random Forest.

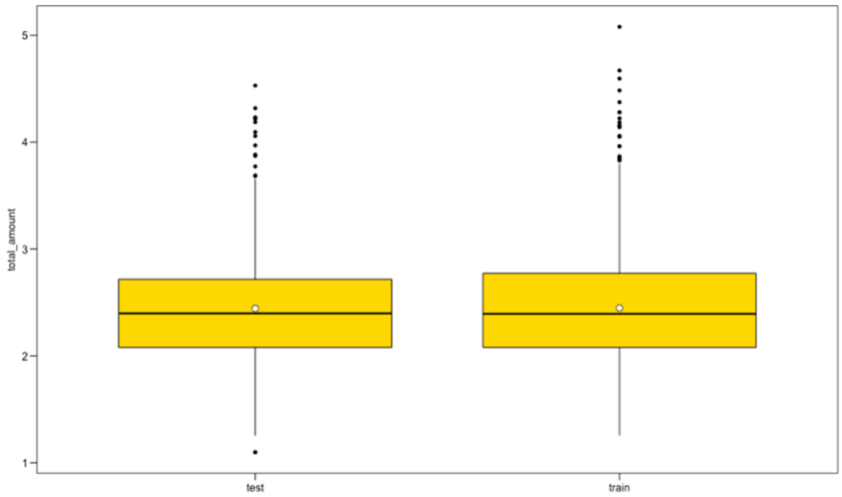

Figure 3: Test vs. train data set comparison

IV. Results and Discussion

Residuals of the OLS and GAM processes were tested for residual normality, with both rejecting the null hypotheses (Table 3). After kriging of the residuals from the GAM was performed, to account for spatial dependence, initial estimates were adjusted with kriging results and after performing the same tests, residuals testing failed to reject the normality hypothesis (Table 3). Final estimates are shown in Table 4. Additionally, the variogram model for the residuals of the GAM is shown in Figure 4. Reduction of the data set down to a smaller subsample, as well as the narrow nature of Manhattan preventing higher horizontal variation in the data, is a possible explanation for the pronounced nugget effect with the cyclical trailing-off pattern in the variogram.

model	statistic	p value	test
OLS	80898	0.000	Jarque Berra
GAM	7195.7	0.013	Jarque Berra
GAM+Krige	1889.1	0.135	Jarque Berra
OLS	0.88984	0.000	Shapiro-Wilk
GAM	0.91051	0.014	Shapiro-Wilk
GAM+Krige	1.0134	0.091	Shapiro-Wilk

Table 3: Tests for residual normality

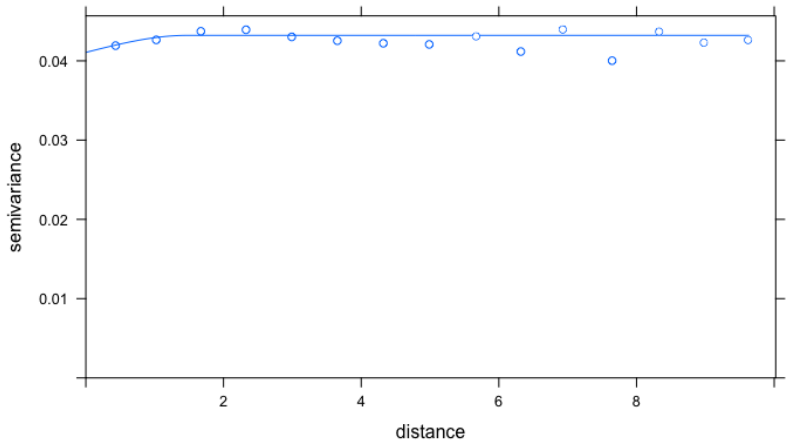

Figure 4: Spherical variogram fit on GAM residuals

Looking at Table 4, having completed all the necessary tests, it can be noted that between two coordinate points, latitude is the relevant one in predicting total wage earned from taxi rides. By no means are these the most important factors, but the original hypothesis remains unrejected with these results. There is a strong case to be made that being uptown versus downtown is what matters more than being on the west or east side of the island. The exact effect of the relationship is quite small in the end, and not worth elaborating on more, as the goal of this project has been accomplished.

	OLS				GAM			
term	estimate	std.error	statistic	p.value	estimate	std.error	statistic	p.value
intercept	-17	66.4	-0.255	0.798	2.38976	0.01407	169.866	0.000
pickup_longitude	1.01	0.678	-0.921	0.357	1.38	1.673	2.744	0.134
pickup_latitude	-0.669	0.519	-1.29	0.185	-0.564	0.321	0.086	0.002
passenger_count	0.00987	0.00588	1.68	0.093	-0.0008	0.0039	-0.193	0.8473
trip_distance	0.148	0.00287	51.5	0.000	7.664	8.511	933.442	0.000

Table 4: Model estimates

Model	RMSE	% Variance Explained
OLS	0.318	71.60
GAM+Krige	0.225	85.1
RandomForest	0.252	81.31

Table 5: Predictive and explanatory power

When it comes to predictions, using the testing subset of the model data, we can see that moving from the OLS, over the GAM, to the Random Forest, there is a decreasing error in being able to predict where highest total wages will be earned (Figure 4; Table 5). It is important to note that the GAM explains most of the variance in the data and has best prediction rates, albeit very close to the Random Forest and kriging combination (see Appendix for visualizations). The expectation for the project was that the Random Forest model would contain significantly better prediction rates. This is important to know because many machine learning models end up being black boxes for the parameters in question, leaving researchers unable to explain great predictions, as well as unable to properly explain the relationships between dependent and independent variables. Kriging is a powerful (and computation heavy) tool for interpolating a random space realization into a usable set of predictions while maintaining the ability to give plausible reasoning behind starting assumptions.

V. Bibliography

- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

- Gámez, M., Montero, J., & Rubio, N. (2000). Kriging methodology for regional economic analysis: Estimating the housing price in Albacete. International Advances in Economic Research, 6, 438–450. https://doi.org/10.1007/BF02294963

- Krige, D. G. (1951). A statistical approach to some mine valuation and allied problems on the Witwatersrand [Thesis]. http://wiredspace.wits.ac.za/handle/10539/17975

- Matheron, G. (1973). The Intrinsic Random Functions and Their Applications. Advances in Applied Probability, 5(3), 439–468. https://doi.org/10.2307/1425829

- Matheron, Georges. (1963). Principles of geostatistics. Economic Geology, 58(8), 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246

- Tin Kam Ho. (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844. https://doi.org/10.1109/34.709601

- Whong, C. (2014, March 18). FOILing NYC's Taxi Trip Data. https://chriswhong.com/open-data/foil_nyc_taxi/

VI. Appendix

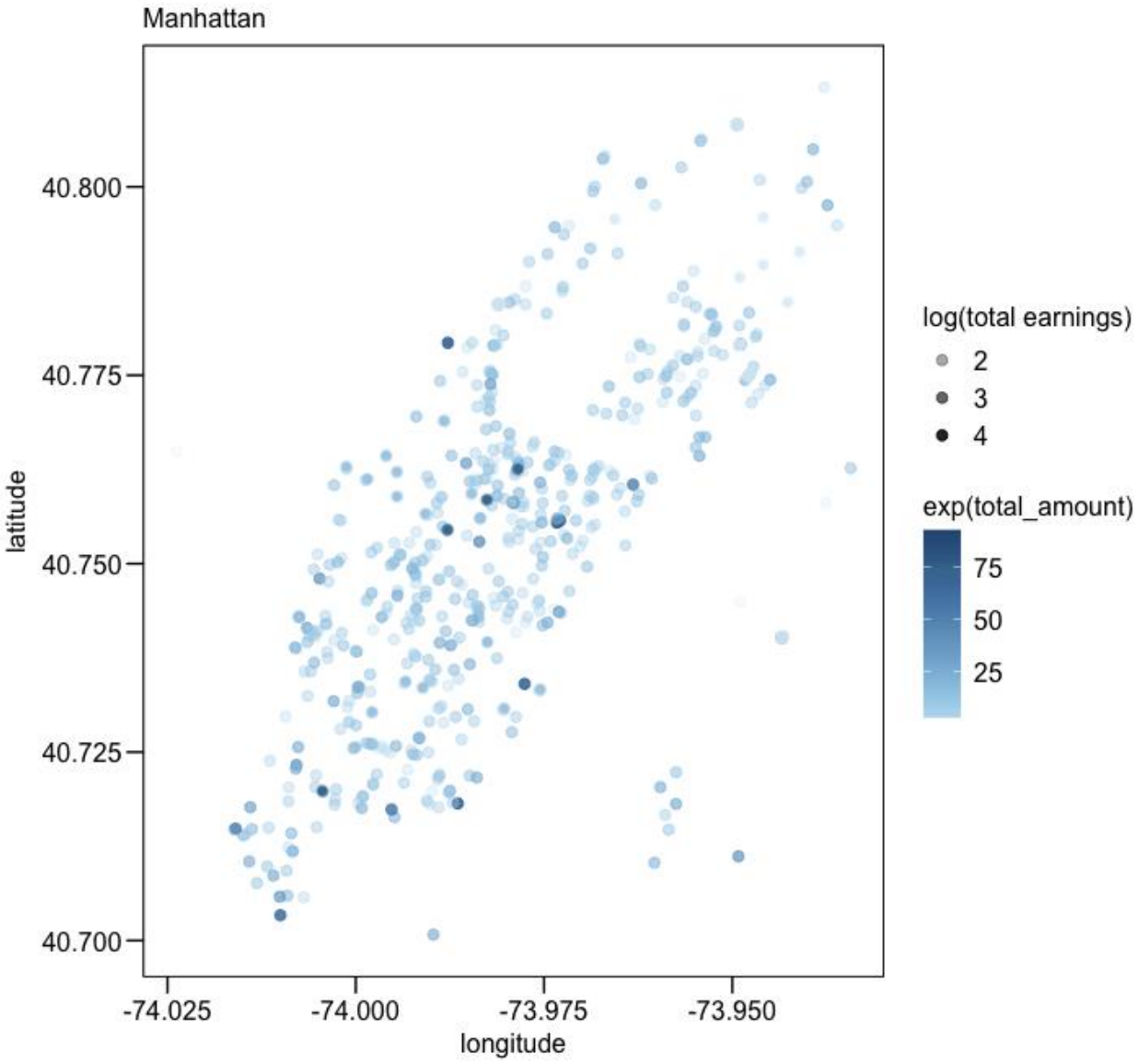

Figure 5: Total earnings from data

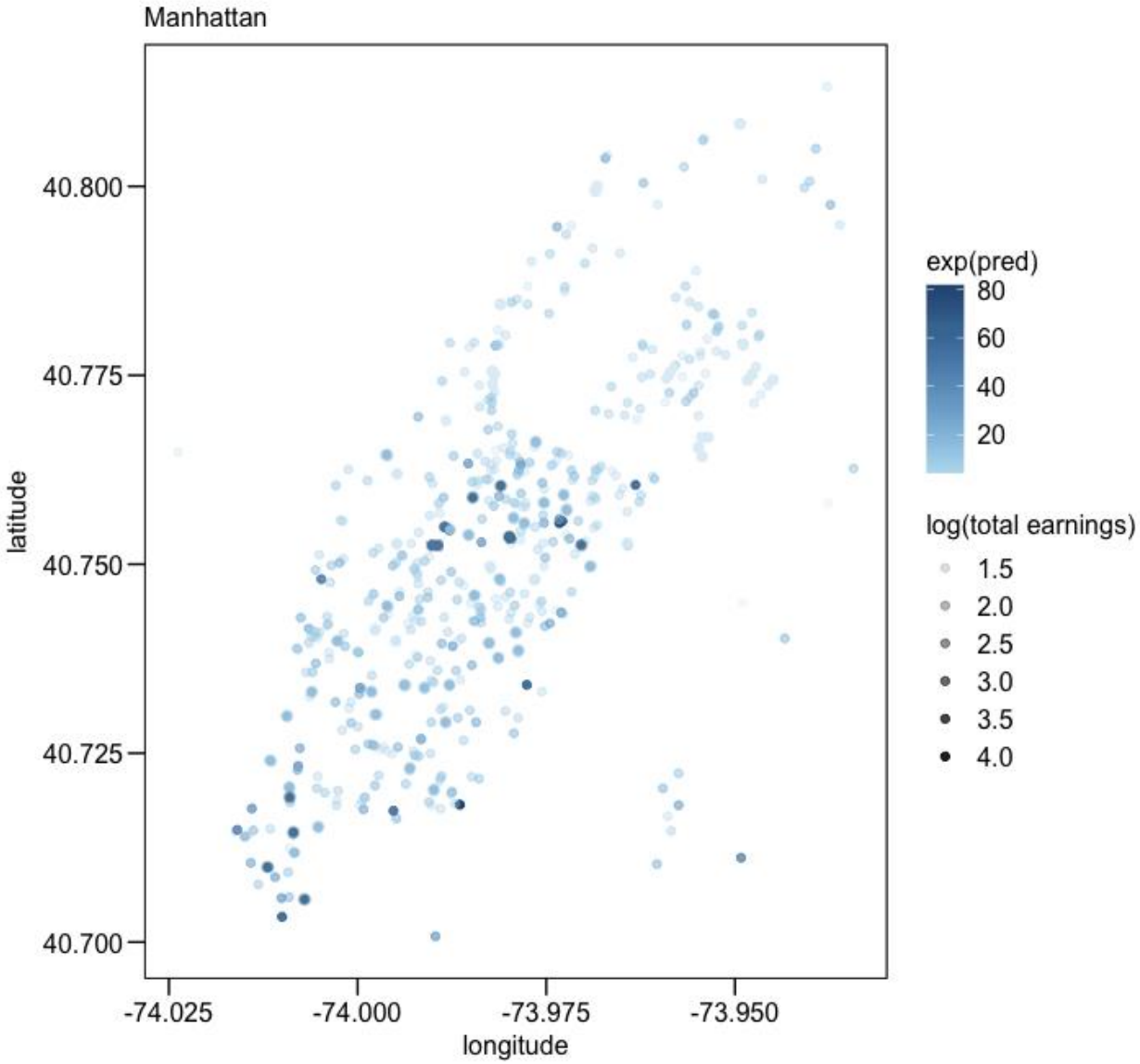

Figure 6: Predicted total earnings from GAM+Krige

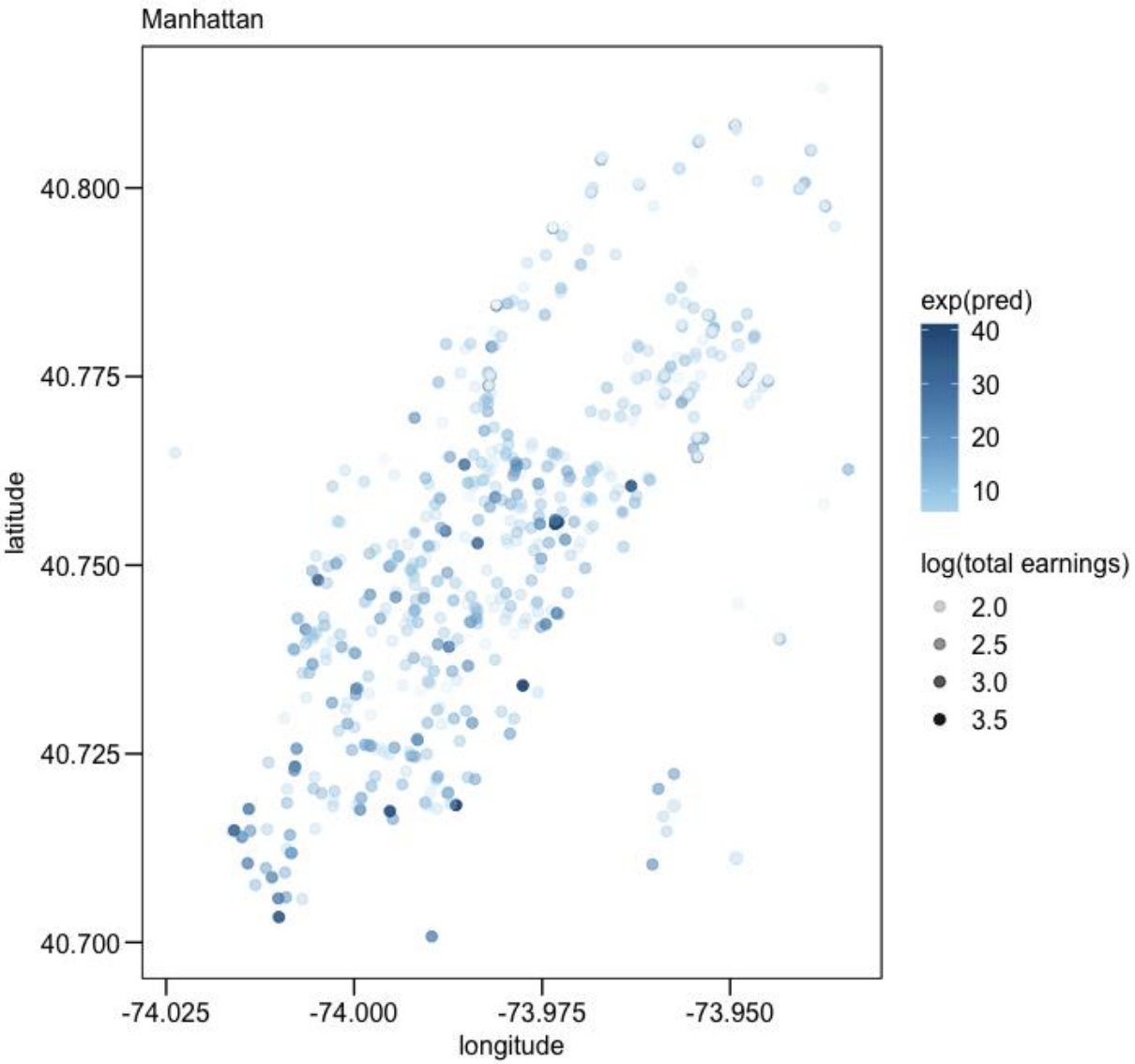

Figure 7: Predicted total earnings from Random Forest

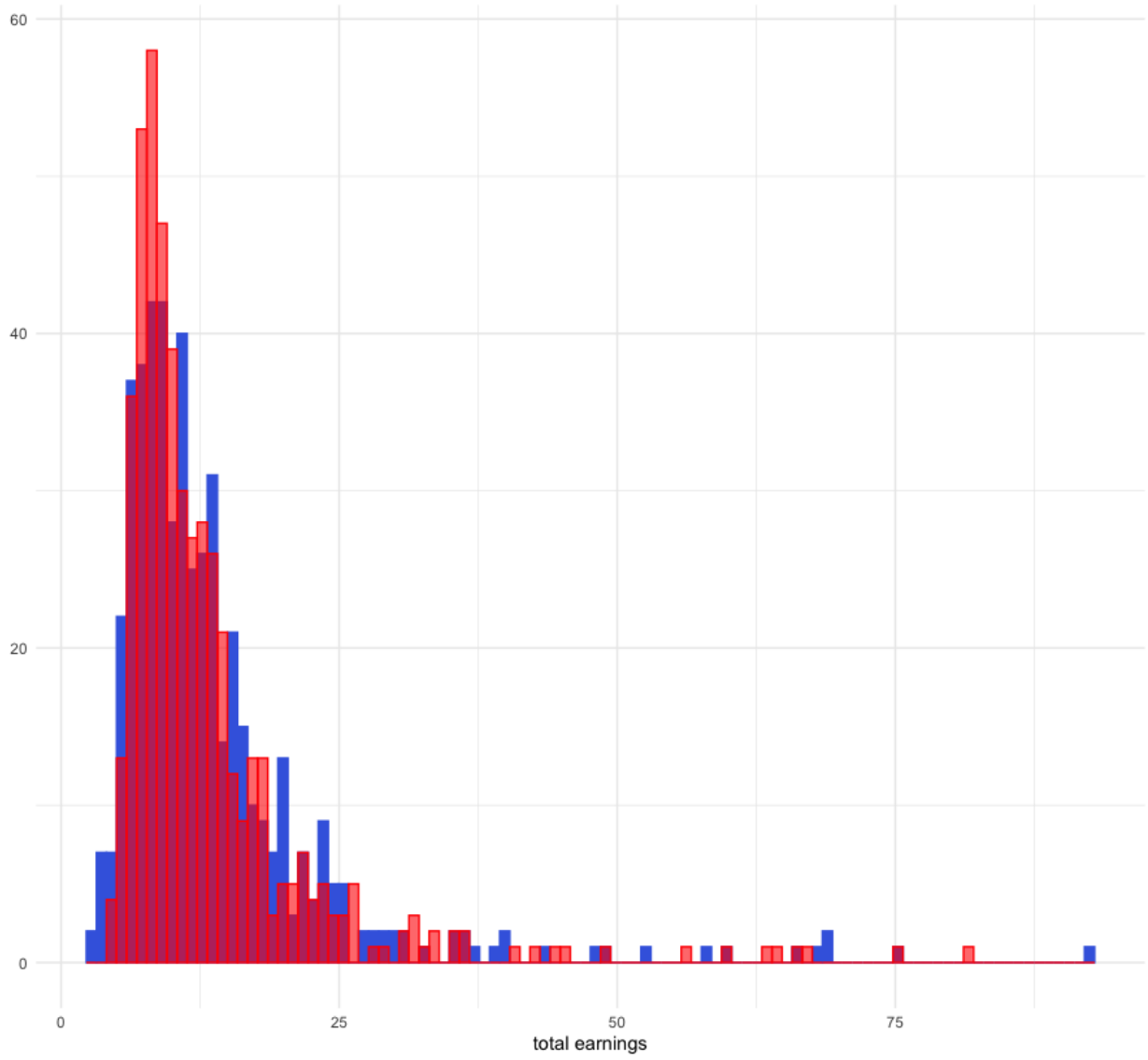

Figure 8: GAM+Krige prediction distribution (red) vs. actual data (blue)