Taxi Usage and City Popularity in Manhattan

Dengyang Xu, Lu Chen, Yijia Zheng, Shen En Chen, Jianing Li, Siyuan Chen

1 MOTIVATION

With the increased coverage of sensors in the cities, researchers now have access to trajectory big data. However, the majority of existing techniques either emphasize statistics or lack context-sensitive interactive visualization capabilities.

2 OBJECTIVE

Our project aims address the shortcomings of previous work by providing context-rich interactive visualizations for analyzing the temporal and spatial relationships between taxi trips in Manhattan in 2021 and the city's popularity.

3 RELATED WORK

The increased coverage of sensors in cities have led to substantial growth in the volume of data and a rapid transformation in the nature of data that records human activity and mobility. Researchers now have access to data that captures high-frequency, real-time, fast-dynamic human activity, which differs dramatically from traditional survey approaches [1].

Since more than fifteen years ago, it has been proposed that GPS technology be utilized to study human mobility and improve transportation surveys [2, 3]. It plays an essential role in transportation modeling and social science applications such as travel mode detection [4], route choice analysis [5], and human mobility behavior analysis [6]. However, given the unstructured nature of big data, trajectory big data must be interpreted within an organized context in order to uncover informative insights and meaningful patterns. Researchers commonly integrate GPS data with highly organized conventional survey data. For instance, trip orientation and end locations were matched with land use data to identify trip intents such as "go home" and "go to work" [7]. Typically, GPS data, GIS data, and demographic data are joined to mine travel data seeking behavior patterns and their correlations with social elements, or infer successive excursions to tour-based corrections [8].

Not only can GPS data help reveal spatial distribution and patterns, but they also allow for a promising analysis of temporal aspects of human mobility. Some researchers have been utilizing the GPS-based data produced by urban infrastructures such as buses, subways, taxis, public utilities, and roads, among which taxi data has attracted considerable interest due to its spatial and temporal granularity and detailed travel information. Several efforts analyzing the taxi data have provided unique insights into different aspects of urban life.

Some research focused on predicting taxi demand, taking into account variables like demographics, land use, transportation accessibility, and weather [9, 10]. The success of deep learning on a variety of computer vision tasks has also prompted academics to adopt the deep neural network for anticipating supply and demand [11, 12]. Some studies dived into the mobility patterns of drivers facing various city environments such as congestion [13]. [14] categorized taxi drivers based on their income and discovered that top-income drivers tend to strike an optimal balance between taxi travel demand and fluid traffic conditions. Besides, other works have centered on human behaviors during taxi rides such as tipping, discovering that factors like average income of pick-up and drop-off sites and weather conditions will affect the likelihood that a passenger tips [15, 16].

Visual analysis tools have been widely adopted to enable interactive and intuitive data exploration. The most common method for taxi data visualization is combining geospatial maps with other graphical representations such as line charts, histograms, and scatterplots to visually capture space, time, and non-spatial temporal characteristics [17]. For example, [18] constructed an interactive system that allows users to visually query taxi trips in New York City and explore regional patterns by applying aggregations and visual results; [19] created a chord diagram plot to visualize Beijing taxi OD flows with different space-time scales; [20] built a node-link graph to represent time-varying transportation networks.

However, most of the existing approaches either **(1) focused on the statistical analysis and hypothesis testing on the correlation between the trajectory data and one particular type of contextual data at a macro view or (2) provided a visualization that purely presented the trajectory information**.

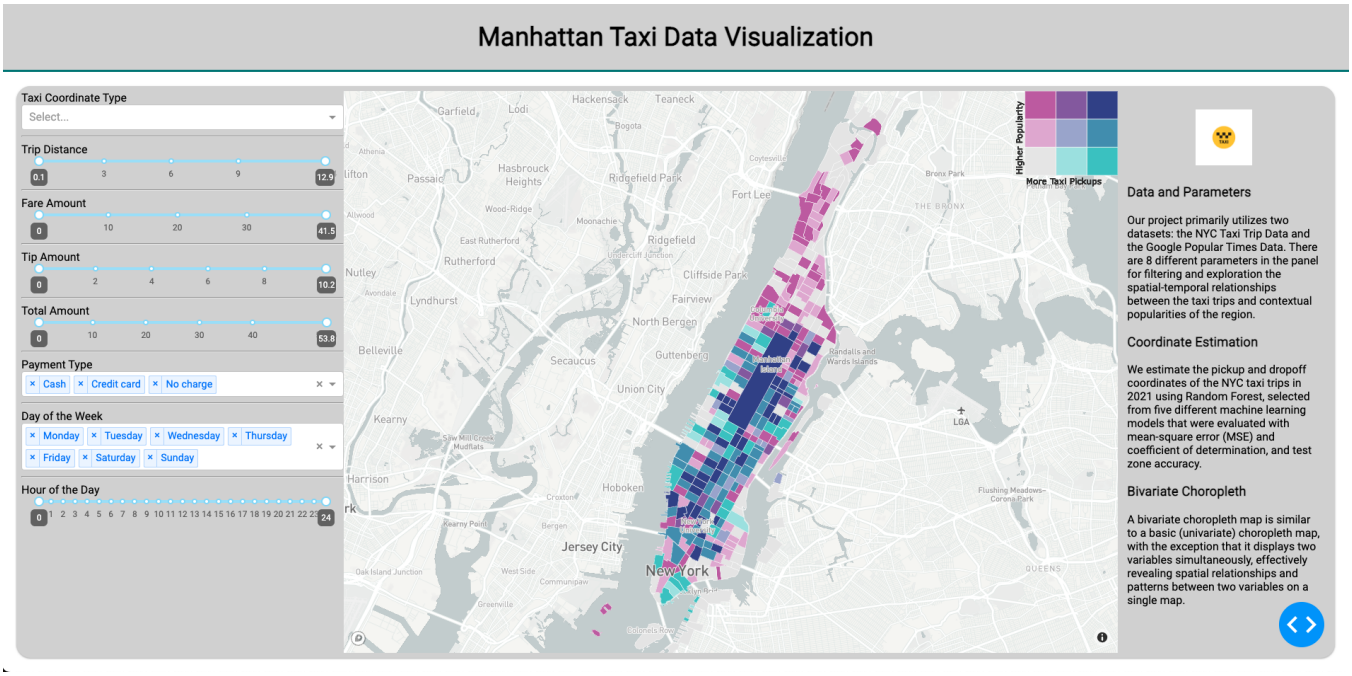

Figure 1: A screenshot of our visualization

There is a lack of an interactive tool that can effectively communicate the temporal and spatial correlations between trajectory data like taxi trips and multiple different contexts so as to provide a more complete picture to understand the dynamics of the city.

4 OUR APPROACH

In this section, we describe the intuition of our work (Subsection 4.1), the dataset used (Subsection 4.2) and the main proposed methods: estimating taxi trip coordinates with classical machine learning models (Subsection 4.3) and using bivariate choropleth to visualize the spatial-temporal relationships between taxi trips and regional popularity (Subsection 4.4).

4.1 Intuition

The intuition of our project is to provide a visualization tool that depicts the spatial-temporal connection of trajectory information and contextual data in a micro perspective, allowing users to analyze the data interactively. This would be a better tool compared to previous work studying taxi trips because past literature often focuses too much on the spatial-temporal characteristics of taxi trips alone and neglect the confounding contextual factors that could impact taxi trip distributions. Further, these analyses often required a certain level of statistical background to understand. By providing a tool that transforms insights into an easy-to-understand interactive visualization, we are reaching out to more audiences that could benefit from the information. Predicting the taxi trip coordinate serves as a necessary step to create our visualization, and hopefully, the results would provide insights for future work on trajectory prediction.

4.2 Dataset

Before we introduce the main two proposed approaches in the following sections, we provide an overview of the dataset used as well as the data cleaning procedures.

4.2.1 Data. To study the **spatial-temporal correlations between taxi trips and regional dynamics**, we obtain records of taxi trips with various attributes about the costs, times, and locations through the 2021 NYC Yellow Taxi Trip Dataset and leverage the Google Popular Times datasets as a proxy of the regional dynamics. We collected the former via NYC Open Data and the latter via the open-source Python API. The taxi trip dataset contains 30.9 million trip records and

the Google Popular Times dataset cover around 30,000 places.

4.2.2 Data Preprocessing. For all of the taxi trip data, we remove records that consist of invalid values. For example, we drop records that have non-positive values for the `passenger_count` attribute. We also remove tuples containing outlier values for any of the attributes. Due to the replacement of coordinate information to taxi zone IDs in the taxi datasets released after July 2016, we have to impute the missing coordinates in order to provide accurate geographical distributions of the trips. As a result, we leverage different classical machine learning models to **estimate the pickup and dropoff coordinates** of taxi trips.

4.3 Coordinate Estimation

We formulate the problem as a multi-output regression problem. For ease of exposition, we consider the case of predicting the dropoff coordinate and fit two regression models, one predicts the latitude and the other predicts the longitude. We performed one-hot encoding on the categorical trip attributes including `pickup_zone` (`dropoff_zone` in the case of prediction pickup coordinates), `ratecodeid`, `payment_type`, `vendor_id`. Then, we analyzed the correlations of each numerical trip attribute such as `trip_distance` and `trip_amount` against the target latitude/longitude to perform simple feature selection. A min-max normalized copy of the training data is also prepared for models that can be affected by the different feature scales.

We originally train the model for 3 million trip data from 2014, which consists of the ground truth coordinates for the models to learn. Within a few iterations, we notice the massive amount of data is costing an unexpected amount of computes; as a result, we reduce the data points to 1000.

We experimented with five classical models: linear regression, decision tree, k-nearest neighbor regression, gradient boosting, and random forest. The 1000 data points from 2014 are randomly sampled from November 2014 and the test data are 1000 trip records from November 2021. Since the Google Popular Times data could change over time and the data we use was collected in November 2022, we believe that the seasonal distribution shifts, if any, would be minimized if we consider data all in the timeframe of November. We split the 1000 data points from 2014 into 800 training data and 200 validation data.

During training, we evaluate the models using mean squared error (MSE) and coefficient of determination (R^2 score) on both the training and validation split. The former is a common non-negative metric used to measure the average squared difference between the estimated values and the actual value; the lower the value the closer the better the model performs. The latter is also known as the "goodness of fit"; it is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor, represented as a value between 0.0 and 1.0. Since the test data has no ground truth coordinates for the pickup and dropoff locations, we calculate the model accuracy on predicting a coordinate in the correct taxi zone, which we refer to as test taxi zone accuracy. We report and discuss the quantitative and qualitative results in Section 5.

4.4 Bivariate Choropleth

Our project primarily utilizes two datasets: the NYC Taxi Data and the Google Popular Times Data. With millions of trip records, the former highlights the geospatial aspects, while the latter stresses the time-varying characteristics of specific locations. Consequently, a bivariate choropleth map is an optimal method for integrating two datasets into a single map, enabling us to depict the spatial-temporal connection of trajectory information and contextual data from a micro perspective. Moreover, because the map encodes two variables simultaneously, it is visually efficient and displays low latency.

A choropleth map is a thematic map in which regions (typically geographic areas, such as boroughs, states, and countries) are colored, shaded, or patterned according to the value of a single variable that corresponds to the area. A bivariate choropleth map is similar to a basic (univariate) choropleth map, with the exception that it displays two variables simultaneously, which has the potential to effectively reveal spatial relationships and patterns between two variables on a single map, such as whether they agree or disagree, increase or decrease proportionally, etc.

Each location on our map corresponds to a census tract in New York City. We designed a 9-class bivariate color scheme by merging two sequential color schemes,

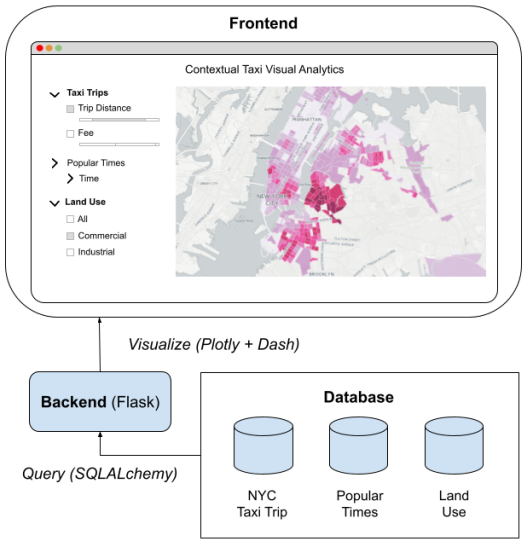

Figure 2: The design of our visualization platform

with the X axis reflecting the number of taxi pickups per area and the Y axis showing the average popularity of the area. The dark blue cell in the upper right corner of this grid reflects cases in which both variables are high, corresponding to the most densely populated area on the map, Central Park.

4.5 User Interface

The visualization tool consists of three main parts: the back-end server, visualization, and front-end interface (Figure 2). We elect `Flask` to be the framework because, unlike the Django framework, Flask is very Pythonic with easy-to-understand APIs. The relatively shallow learning curve of it allows us to more equally distribute development among the member of different coding experiences. We use `SQLALchemy` API to communicate with the data sources. For the visualization part, we utilize `Plotly` to create bivariate choropleths based on the data drawn from the back end. As for the front-end interface, `Dash` is a python framework created by `Plotly` for creating interactive web applications. The dashboard contains seven filters: Coordinate type, Trip Distance, Fare Amount, Tip Amount, Total Amount, Payment Type, Day of the Week, and Hour of the Day. The users can filter and customize the data based on these parameters to view Manhattan's taxi popularity and pick-up information in advance.

5 EXPERIMENT EVALUATION

In this section, explain the details of our experiments. In Section 5.1, we provide a list of motivating questions our work attempts to answer. In Section , we analyze the quantitative results and discuss the qualitative observations of the coordinate estimation described in Subsection 4.3. In Subsection 5.3, we detailed our attempts in visualizing the data and observations from the bivariate choropleth.

5.1 Motivating Questions

At the core, estimating the coordinate provides us with more coherent and accurate geospatial information about taxi trips when we combine them with the Popular Times to analyze regional dynamics. Through our experiments, we hope to identify the best-performing model to achieve this objective and analyze the reasons for different model performances. Only after we obtain more recent coordinate information could we analyze the spatial-temporal relationships between taxi usage and city popularity.

5.2 Coordinate Estimation

5.2.1 Quantitative Results and Analysis. For coordinate estimation, we have detailed the experiment setup in Subsection 4.3. For each of the five different methods, we train it 10 times and report the best performance in Table 1. As we can see, the random forest model performs the best, followed by linear regression with comparable numbers for all metrics. Decision trees perform worse than the random forest, which is expected, as we learn in class that the bootstrapping and aggregation components of random forests lead to a lower variance model when compared to a single decision tree. The K-nearest neighbor model performs very poorly; more than 96% of the predictions are coordinates that are outside of their ground truth taxi zones. This is expected because there could be data shifts in the feature space between 2014 and 2021, and k-nearest neighbor regression does not take such shifts into account when making predictions based on the neighbors.

5.2.2 Qualitative Observation. We visualize the predictions of each model in Figure 3, Figure 4, Figure 5, Figure 6, and Figure 7. Qualitatively, the predictions of the random forest and linear regression models (Figure 3, and Figure 5) form the most natural scatter in

Model	Train MSE	Validation MSE	Train R^2	Validation R^2	Test Zone Accuracy
Random Forest	$6.36 \cdot 10^{-6}$	$1.22 \cdot 10^{-5}$	0.99	0.97	0.949
Gradient Boost	$7.97 \cdot 10^{-5}$	$8.52 \cdot 10^{-5}$	0.82	0.77	0.095
Linear Regression	$7.13 \cdot 10^{-6}$	$1.05 \cdot 10^{-5}$	0.98	0.97	0.930
KNN Regression	$3.32 \cdot 10^{-4}$	$4.65 \cdot 10^{-4}$	0.22	-0.27	0.035
Decision Tree	$3.33 \cdot 10^{-6}$	$1.79 \cdot 10^{-5}$	0.99	0.95	0.737

Table 1: Quantitative results of coordinate prediction

Manhattan. For the predictions of the decision tree (Figure 7), we can see a clear vertical line of scatters aligned in the center of the figure, indicating the predictions are snapped to some implicit grids. The scatters of the gradient boost model form an infinity sign and a vertical line of scatters like that for the decision tree is also observed (Figure 4). Lastly, the k-nearest neighbor regression yields the most noticeably poor performance, with most of the scatters clustered around the south end of Central Park (Figure 6).

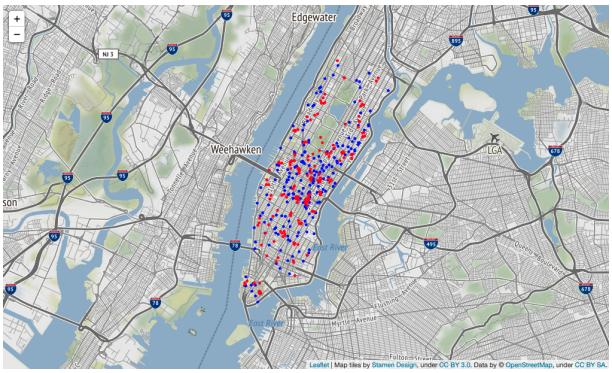

Figure 3: Predictions (red) and ground truths (blue) of the random forest model

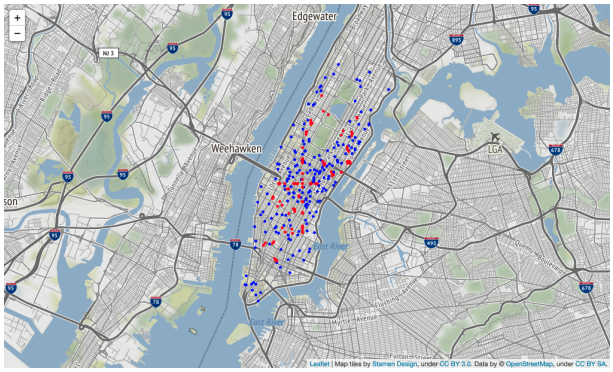

Figure 4: Predictions (red) and ground truths (blue) of the gradient boost model

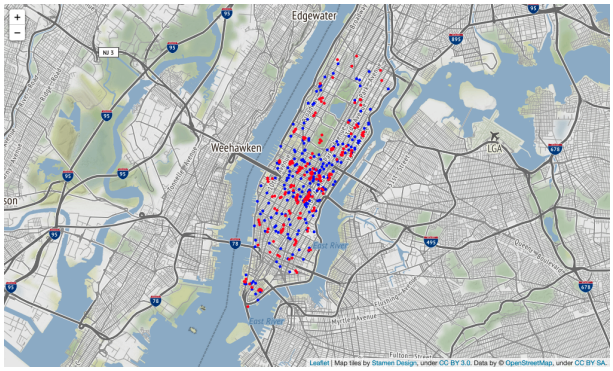

Figure 5: Predictions (red) and ground truths (blue) of the linear regression model

5.3 Bivariate Choropleth

5.3.1 Previous Attempts. While we propose the bivariate choropleth as the final form of visualization, we experimented with many different visualizations to effectively present data of two distinct spatial-temporal characteristics. Over the course of the semester, we prototyped density heat maps, bubble maps, and Kringing interpolation [21]. The density heat maps are effective in capturing the geospatial variability of taxi trip locations and the bubble maps are useful for highlighting the popularity of different locations at a given time. Kringing interpolation is a geostatistical technique based on statistical models that include autocorrelation, producing a prediction surface as well as a measure of the certainty or precision of the predictions. It stretches the Popular

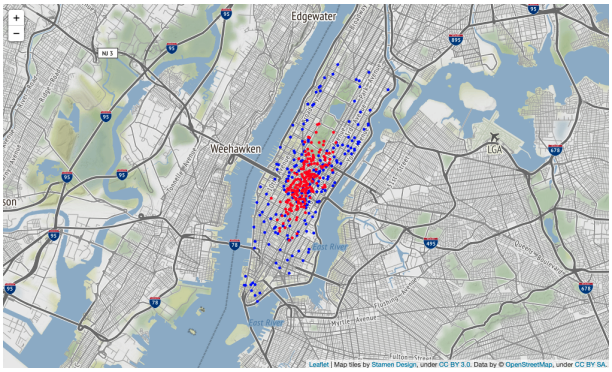

Figure 6: Predictions (red) and ground truths (blue) of the k-nearest neighbor regression model

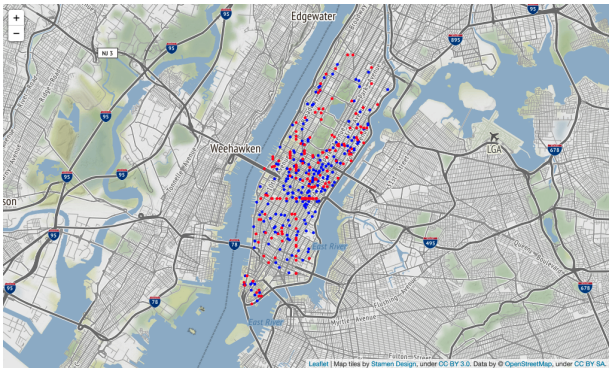

Figure 7: Predictions (red) and ground truths (blue) of the decision tree model

Times data from limited points of interest (POIs) to areas in between with a smooth surface connecting the points. These methods, however, turn out to have shortcomings. The density heat maps and bubble maps can highlight well the properties of the corresponding data, but, when combined, a diminishing effect is observed due to the lack of coherence between the two types of visualization. Kringing interpolation requires a large number of data points and computations.

5.3.2 Observations. After we finalize our visualization with a bivariate choropleth, we uncovered several interesting facts:

(1) On average, Central Park is the most popular region in Manhattan, reflected by high popularity and a high number of taxi pick-ups and drop-offs. Interestingly, the northern part of Manhattan has similar popularity as Central Park but only a few taxi trips to or from the area.
(2) There are more short-distance pick-ups and drop-offs around Central Park, and the fare per mile is the highest. We hypothesize that it is because the heavy traffic and frequent stopping will result in a surcharge, especially during rush hours.
(3) The tip mount near Central Park is the most expensive, which is expected since it is the most touristy area with likely higher prices of goods.
(4) Credit card is the primary payment method for taxi rides with a predominantly high proportion among all rides.

6 DISCUSSION

The bivariate choropleth unveils a number of intriguing characteristics of taxi usage and city popularity. The different filters provided on the interface enable us to facet the data in almost any way we could possibly think of when exploring the data. While rather simple, predicting the taxi trip coordinates with classical machine learning models also allows us to combine the data fairly in terms of their time frames. While the bivariate choropleth effectively resolves the problem of heterogenous spatial-temporal properties of the two data sources, there are still opportunities for improvement. Due to computational constraints and the requirements to visualize in interactive time, we had to down-sample the datasets and limit the points to Manhattan. This, unfortunately, prevents us from analyzing the dynamics of the entire city of New York.

7 CONCLUSION

Despite the shortcomings discussed in Section 6, we believe our bivariate choropleth effectively and efficiently illustrates the dynamics between taxi usage and regional popularity for Manhattan. In future work, we plan to scale our visualization to the scope of NYC by leveraging more efficient indexing and caching of the data to achieve faster filtering and retrieval.

8 CONTRIBUTIONS

All team members have contributed a similar amount of effort.

REFERENCES

[1] Michael Batty. Big data and the city. *Built Environment*, 42(3):321–337, 2016.

[2] Jesse Casas and CH Arce. Trip reporting in household travel diaries: A comparison to gps-collected data. In *78th annual meeting of the Transportation Research Board, Washington, DC*, volume 428, 1999.

[3] Jean Wolf, Marcelo Oliveira, and Miriam Thompson. The impact of trip underreporting on vmt and travel time estimates: preliminary findings from the california statewide household travel survey gps study. *Transportation Research Record*, 1854:189–198, 2003.

[4] Peter Widhalm, Philippe Nitsche, and Norbert Brändie. Transport mode detection with realistic smartphone sensor data. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 573–576. IEEE, 2012.

[5] Jeffrey Hood, Elizabeth Sall, and Billy Charlton. A gps-based bicycle route choice model for san francisco, california. *Transportation letters*, 3(1):63–75, 2011.

[6] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[7] Jean Wolf, Randall Guensler, and William Bachman. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record*, 1768(1):125–134, 2001.

[8] Li Shen and Peter R Stopher. A process for trip purpose imputation from global positioning system data. *Transportation Research Part C: Emerging Technologies*, 36:261–267, 2013.

[9] Ci Yang and Eric J Gonzales. Modeling taxi trip demand by time of day in new york city. *Transportation Research Record*, 2429(1):110–120, 2014.

[10] Annick Lacombe and Catherine Morency. Modeling taxi trip generation using gps data: the montreal case. Technical report, 2016.

[11] Dong Wang, Wei Cao, Jian Li, and Jieping Ye. Deepsd: Supply-demand prediction for online car-hailing services using deep neural networks. In *2017 IEEE 33rd international conference on data engineering (ICDE)*, pages 243–254. IEEE, 2017.

[12] Filipe Rodrigues, Ioulia Markou, and Francisco C Pereira. Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. *Information Fusion*, 49:120–129, 2019.

[13] Liang Liu, Clio Andris, and Carlo Ratti. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6):541–548, 2010.

[14] Marco Veloso, Santi Phithakkitnukoon, Carlos Bento, Nuno Fonseca, and Patrick Olivier. Exploratory study of urban flow using taxi traces. In *First Workshop on Pervasive Urban Applications (PURBA) in conjunction with Pervasive Computing, San Francisco, California, USA*, 2011.

[15] David Elliott, Marcello Tomasini, Marcos Oliveira, and Ronaldo Menezes. Tippers and stiffers: An analysis of tipping behavior in taxi trips. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8. IEEE, 2017.

[16] Won Kyung Lee and So Young Sohn. A large-scale data-based investigation on the relationship between bad weather and taxi tipping. *Journal of environmental psychology*, 70:101458, 2020.

[17] Antoine Clarinval and Bruno Dumas. Intra-city traffic data visualization: a systematic literature review. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[18] Nivan Ferreira, Jorge Poco, Huy T Vo, Juliana Freire, and Cláudio T Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE transactions on visualization and computer graphics*, 19(12):2149–2158, 2013.

[19] Huihui Wang, Hong Huang, Xiaoyong Ni, and Weihua Zeng. Revealing spatial-temporal characteristics and patterns of urban travel: A large-scale analysis and visualization study with taxi gps data. *ISPRS International Journal of Geo-Information*, 8(6):257, 2019.

[20] Xiaoke Huang, Ye Zhao, Chao Ma, Jing Yang, Xinyue Ye, and Chong Zhang. Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE transactions on visualization and computer graphics*, 22(1):160–169, 2015.

[21] Zoubeida Kebaili Bargaoui and Afef Chebbi. Comparison of two kriging interpolation methods applied to spatiotemporal rainfall. *Journal of Hydrology*, 365(1-2):56–73, 2009.