|  |  |
| --- | --- |
| 1 | |
| Paper Name | Urban Mobility Study using Taxi Traces |
| Author | Marco Veloso, Carlos Bento and Santi Phithakkitnukoon |
| Publishing Year | 2011 |
| Case Study | Lisbon, Portugal |
| Dataset | * 10 million taxi-GPS samples from August through December in 2009, collected in Lisbon, Portugal by GeoTaxi * Lisbon council and population density map/count * Sapo Maps provided a collection of 10,954 Points of Interest (POIs), grouped into eight categories (Services, Recreation, Education, Shopping, Police, Health facilities, Transportation and Accommodation), to characterize the area type. * Weather conditions for the period under study were retrieved from Weather Underground [11] and grouped in three states (sunny, cloudy and rainy). |
| Aim of the paper (Abstract) | * Perform an exploratory analysis to visualize the spatiotemporal variation of taxi services * Explore the relationships between pick-up and drop-off locations * Analyze the behavior in downtime (between the previous drop-off and the following pick-up) * Carry out the analysis of predictability of taxi trips for the next pick-up area type given history of taxi flow in time and space |
| Target Users | Understand what drives the common citizen, what their needs are. |
| Main Body / Methods | * Lisbon map was divided into 0.5X0.5 km2 grid * Analyzed the population density (grid wise) with the 8 POIs to understand how population is spread out with respect to the predominant POI categories in each location * Exploratory analysis to identify patterns and variables that model the pattern:   + Spatial and Temporal distribution- Hourly and day of the week bar plots   + Grid-wise taxi distribution pattern based on PU and DO locations   + Analysis how strongly connected locations are according to taxi services- Strong relations can be observed between all those locations are characterized by some public transportation modality (airport, train, ferry, bus).   + Relation between pick-ups and drop-off considering only the most frequent destination for each location.   + Taxi service distribution according to distance (top), duration (middle) and income (bottom).   + Fitted the trips distance with a gamma distribution (need to understand this)   + Variation in service distance and income * Downtime analysis- what happens in between services (i.e., downtime – time spent looking for next pick-up):   + Spatial distribution according to the average distance traveled during downtime and the relationship between previous drop-off and next pick-up location - The areas away from the city center (characterized by a higher number of residential buildings) show higher average distances traveled between services, whereas in downtown the distances traveled are relatively smaller   + Average downtime and distance traveled * Predictability Analysis (to predict the taxi movement based on the last drop-off location history):   + Exploratory study shows the possibility of some movement patterns   + Applies a naïve Bayesian classifier for the study. The classifier simply applies the Bayes’ theorem with independence assumption. The objective is to compute the likelihood of each possible pick-up area type given the hour of the day, day of the week, weather condition and area type of the last drop-off.   + Based on 10-folds cross validation, they were able to predict (for each drop-off) the next pick-up area type at about 54% |
| Results |  |
| Conclusion | Being able to accurately predict taxi flow is important and a challenging problem, which we will address it further in our future work. |
| Way Forward (if any) |  |
| Key Takeaways | We can explore the Predictability analysis further |

|  |  |
| --- | --- |
| 2 | |
| Paper Name | Visualization Tool for Taxi Usage Analysis: A case study of Lisbon, Portugal |
| Author | Postsavee Prommaharaj, Marco Veloso, Carlos Bento and Santi Phithakkitnukoon |
| Publishing Year | 2016 |
| Case Study | Lisbon, Portugal |
| Dataset | One week (1-7 June 2009) taxi information for demo purpose |
| Aim of the paper (Abstract) | Develop a visualization tool for exploratory analysis of taxi usage and behavior (Mobility and Flow). It is aimed at providing the first-hand information for the users. |
| Target Users | Useful for taxi service providers in scheduling and dispatching management, as well as urban planning and design |
| Main Body / Methods | Developed an interactive visualization interface for users to understand the taxi movement patterns within the city. Visualization had two modes:   * Mobility-It gives the overview of the statistical inferences of the taxi movements such as PU and DO locations, taxi availability status, number of trips (hourly, weekly etc.), ratio of taxi status etc. Individual taxi level information can be inferred in this mode. * Flow- It provides the spatial travel pattern of taxi such as fireball PU and DO travel flow pattern, location where the status of taxi changed (available and occupied) etc.   User experience survey study was conducted among 50 participants. |
| Results | Tool was mostly found to be enjoyable and easy to start using. |
| Conclusion | They believe that there are several stakeholders for our developed tool that include taxi service providers, transport engineers, and urban planners. |
| Way Forward (if any) | The results also suggest that we need to improve the easy-to-use aspect of our visualization |
| Key Takeaways | The paper had great visualizations which can enhance the data interpretability and explain the technical statistical inferences to policy makers or other stakeholders in a much simpler format.  We can take visualization inspirations from this paper. |

|  |  |
| --- | --- |
| 3 | |
| Paper Name | Impact of weather on urban transit ridership |
| Author | Abhishek Singhal, Camille Kamga, Anil Yazici |
| Publishing Year | 2014 |
| Case Study | New York City |
| Dataset | Two years of ridership data from the MTA-NYCT and 30-year normal hourly and daily temperature weather data from National Oceanic and Atmospheric Administration (NOAA) and Weather Underground website |
| Aim of the paper (Abstract) | This paper utilizes hourly ridership and hourly weather data to model the weather effects. Secondly, the paper developed separate regression models for elevated and non-elevated stations and investigate the weather impacts. Additionally, we investigate the role of connecting bus service routes in maintaining stable station ridership under adverse weather conditions. |
| Target Users | Useful for taxi service providers in scheduling and dispatching management, as well as urban planning and design |
| Main Body / Methods | * They took residual ridership as the dependent variable which means the percentage difference between the actual daily ridership, rt, and an average ridership variable. * Their independent variable is defined in the analysis using temperature, wind speed, wind gusts, rain, snow, and fog information in the weather data. * Apart from weather related variables, fall, winter and spring seasons were included as dummy variables since the effect of weather elements vary by season with respect to summer (base) season |
| Results | * Temporal analysis of weather impact based on hourly and daily ridership- Time-of-day models provide indirect hint about weather impact with respect to trip purpose. Weekend trips (which are discretionary in nature) are more likely to be affected by inclement weather than weekday trips * Daily ridership versus hourly ridership models- For both weekdays and weekends models, the sign of beta coefficients is consistent when directly comparable. On one hand, both daily models have higher adjusted R-square as compared to hourly models. The set of significant variables are different between daily and hourly models. For instance, snow and hot day conditions are significant in daily ridership model but are absent in the hourly model. * Analysis based on time-of-day models- Considering the time-of-day models, over weekends, rain has a negative impact for all time periods however midday and PM ridership are more severely affected than AM ridership. This suggests that for subway riders, future increase in rain events in New York City would more likely affect morning ridership for daily commuters (due to weekday AM travel) and midday and PM ridership for discretionary riders (over weekends). |
| Conclusion | * The time-of-day models indicate that under the given weather conditions, for any day of week, the ridership during the PM time period is most affected, followed by midday period and least affected during AM period. * Time of day-based policy measures like higher service frequencies for the affected time periods and increasing the trip transfer duration may allow more riders to embrace transit service under adverse weather conditions. * The analysis reveals that hourly ridership models are better at including the effect of individual weather conditions and there are more weather conditions that affect the travel behavior of transit riders than those suggested by daily models. |
| Way Forward |  |
| Key Takeaways | The paper takes a different approach of measuring the impact of weather at daily and hourly ridership level. This reveals an interesting observation how different seasons, weather conditions have different effect on ridership count at different time of the day and day of the week. This approach will be very interesting to adopt for our project. |

|  |  |
| --- | --- |
| 4 | |
| Paper Name | Uber vs. Taxis: Event detection and differentiation in New York City |
| Author | Grant McKenzie, Carlos Baez |
| Publishing Year | 2016 |
| Case Study | New York City |
| Dataset | Total, 80,295,320 Yellow taxi pick-up locations and 4,534,327 Uber[[1]](#footnote-1) pick-up locations between April 1 and September 30, 2014 |
| Aim of the paper (Abstract) | This paper uses a sample of Uber and taxi pick-up times and locations in New York City to show that events can be detected within each platform. Additionally, showing that there is a difference in the types of events that are attended by Uber users and taxi passengers. |
| Target Users |  |
| Main Body / Methods | * The timestamps for each pick-up were rounded to the nearest hour and aggregated to counts by intersecting with the New York city census tract spatial data from 2014. * Events were detected in each dataset by comparing the number of pick-ups on any given day, time and census tract with the number of pick-ups typical for that hour of the week (mean count). * An event was recorded if the number of pick-ups was above three standard deviations from the mean. |
| Results | * 485 events were discovered in the Uber dataset and 2,671 in the taxi data. * 17 events identified in both the taxi and Uber datasets * This suggests that certain types of events are aligned with Uber users while others lend themselves to taxi passengers manually investigated several the events based on their location and temporal parameters. * Using application programming interfaces (APIs) such as Eventful 4 along with venue specific websites (e.g., Madison Square Gardens), we were able to identify several events |
| Conclusion | While some events were identified in both the taxi and Uber data, a larger number were detected within only one of the datasets. |
| Way Forward | * Explore event detection at a variety of spatial and temporal resolutions. * The time-window during which events occur ranges significantly depending on the type of event. This will be explored in greater detail along with any demographic information associated with the individuals that attend these events |
| Key Takeaways | The paper demonstrates a possible approach on using taxi data to detect events. The anomaly detection in ridership due to events can be used in predicting taxi demands for future events of similar scale. |

|  |  |
| --- | --- |
| 5 | |
| Paper Name | Exploring the Relationship between Mobile Phone Call Intensity and Taxi Volume in Urban Area |
| Author | Marco Veloso, Santi Phithakkitnukoon, and Carlos Bento |
| Publishing Year | 2012 |
| Case Study | Lisbon, Portugal |
| Dataset | * 500,000 taxi location points collected from 230 taxis through December in 2009 (period of 31 days), collected in Lisbon, Portugal by GeoTaxi (20% market share). Data about GPS location, speed, bearing, engine status, and occupancy status of the taxi. * Speech traffic for each cell site aggregated hourly from TMN (40% market share). The number of calls for each cell site is described as call intensity. * Lisbon council map |
| Aim of the paper (Abstract) | * Perform an analysis of mobile phone call intensity and taxi volume in Lisbon Portugal * Discover the inter-predictability between phone network flow and taxi traffic network flow |
| Target Users | How can we improve our understanding of urban settings by leveraging new tracking technologies (understand the inter-predictability of mobile phone and taxi traffic data) |
| Main Body / Methods | * Lisbon map was divided into 0.5X0.5 km2 grid cells * Exploratory analysis to identify patterns and variables that model the pattern:   + Spatial distribution of taxi volume (grid wise) and try to detect possible hotspots (usually near airports, city center, etc.)   + Line chart comparing temporal distribution of mobile phone call activity and taxi volume   + Quantify the difference between these two time series, we computed the Euclidean distance (ED) for both the whole day and active period (8AM – 11PM)   + Also compare weekdays and weekends time series and ED * Predictability Analysis (to predict the taxi volumed based on phone call intensity and the other way around):   + Searching for highest possible correlation by shifting the two time series between -5 and +5 hours looking for the highest possible value of , focusing one the amount of variance explained by the model (linear regression)   + Line chart representing ED and for different hourly shifts of the two time series |
| Results | * Phone call intensity and taxi volume follow similar patterns, increase around 7AM and drop down after 7PM * Reduction of the activity from both services on weekends and holidays. * Euclidean distance (ED) of the two time series turned out to be 0.21 (general) and 0.1917 in active periods. Euclidian distances lower in weekdays (more activities). * The variation in the amount of taxis is an indicative variable for the mobile phone call intensive of the next two hours. From the fitted regression model, the resulting = 0.8512. * Taxi volume is a predictor of phone call intensity in AM hours, while phone call intensity is a predictor of taxi volume in PM hours |
| Conclusion | Studies of urban settings are usually too simplistic. In order to predict taxi volume looking at just phone call intensity is not enough. |
| Way Forward (if any) | In this case, the authors analyzed patterns and time series over a single month. A temporally wider analysis could unveil more inter-predictability between taxi and phone calls data. Also, phone call intensity might not be a relevant indicator in our world (maybe tweets, location tags, etc.) |
| Key takeaways from the paper | The breakthrough of tracking technologies allows to measure any kind of interaction in the urban setting. The way in which sensor data can be leveraged is up to the researcher. |

|  |  |
| --- | --- |
| 6 | |
| Paper Name | Exploring the Relationship between Mobile Phone Call Intensity and Taxi Volume in Urban Area |
| Author | Zhizhen Liu, Hong Chen, Yan Li, and Qi Zhang |
| Publishing Year | 2020 |
| Case Study | Xi’an, China |
| Dataset | * GPS Data: from the Xi’an Taxi Management Office and consist of vehicle location data that are recorded every 5 s for 30 days. The dataset consists of 40 million track points * Environmental Data:   + Air quality data: pollutant data including PM2.5 and PM10 from the official website of Green Breathing (seven dimensions)   + Meteorological data: from the National Meteorological Information Center. This study selects the hourly data of Xi’an with temperature, humidity, etc. (five dimensions). * Map of Xi’an |
| Aim of the paper (Abstract) | * Predict the demand for taxis in hotspots by constructing a set of affecting factors of the taxi demand for April 2017 (30 days) * Detect hotspots and use GPS and environmental data to predict taxi demand with three different models (random forest, ridge regression, combination forecasting) for April 2017 |
| Target Users | Helping agencies managing taxis to allocate optimal demand in hotspots. |
| Main Body / Methods | * Models:   + Random Forest Model (RFM): ensemble learning method of learning where at each node of each decision tree a subset of k random features are selected to be used for the split decision.   + Ridge Regression Model (RRM): alternative version of the OLS (losing unbiasedness) that fits the ill-conditioned data more accurately than the OLS. Introduce the norm regularization term that helps when there are many features and the sample is small. has usually lower but higher significance of estimators.   + Combination Forecasting Model (CFM): weighted predictor of RFM and RRM. The crucial point is the choice of the weight coefficients. * Data processing:   + GPS Data Processing: describes the driving state, passenger or driving state.   + Feature Selection: have to make sure that features are independent or account for that in order to avoid multicollinearity. To find out correlation use the Pearson coefficient, which is equal to the covariance of the variables divided by the product of their standard deviations.   + One-Hot Encoding: turn categorical features into dummies for each category. Now the feature dimension is 39, the first 23 days are used for training and 7 for the testing. |
| Results | * Extract Hotspots: kernel density model identifies transportation hubs, commercial areas. For this study, Bell Tower and Xi’anbei Railway Station are chosen. * RFM prediction: the most influent parameter is the number of estimators. Accuracy can reach 86% for the railway. Feature importance (through VIM) highlights the predominance of first hours dummy variables for Railway and Ozone for Bell Tower. (Railway) = 0.854 * RRM prediction: tune the regularization intensity and the most important feature for both area re the dummies for the hours of the day. (Railway) = 0.864 * CFM prediction: 80/20 weights for RFM and RRM. (Railway) = 0.885 |
| Conclusion | The time factor has the most important effect on taxi demand. However, ozone and temperature variables demonstrated to be important in some settings (Bell Tower area). The proposed CFM is the most accurate one on the test data. |
| Way Forward (if any) | * The impact of similar types of traffic demand is ignored, taxi demand will be reduced if another service takes place (e.g. Uber). * Extend the study to also consider land use properties |
| Key takeaways from the paper | There a myriad of possible datasets related to citizens interactions in a city that can be used to find inter-predictability with transportation modes as (e.g. taxi) |

|  |  |
| --- | --- |
| 7 | |
| Paper Name | Hunting or Waiting? Discovering Passenger-Finding Strategies from a Large-scale Real-world Taxi Dataset |
| Author | Bin Li , Daqing Zhang, Lin Sun, Chao Chen, Shijian Li, Guande Qi, and Qiang Yang |
| Publishing Year | 2011 |
| Case Study | Hangzhou, China |
| Dataset | * Taxi GPS dataset of more than 5350 taxis (200b records) served in Hangzhou for one year (Apr 2009 - Mar 2010) with the following variables:   + Vehicle ID   + Longitude   + Latitude   + Speed   + State   + Timestamp * Hangzhou map |
| Aim of the paper (Abstract) | * Predict the demand for taxis in hotspots by constructing a set of affecting factors of the taxi demand for April 2017 (30 days) * Detect hotspots and use GPS and environmental data to predict taxi demand with three different models (random forest, ridge regression, combination forecasting) for April 2017 |
| Target Users | Helping agencies managing taxis to allocate optimal demand in hotspots. |
| Main Body / Methods | * Data processing and Extraction:   + Reduce the computational burden by considering only 15 working days in Oct 2019 and consider only the metropolitan area.   + For each PU/DO event, we extract the GPS data and event timestamp. Also we consider variables that allow to understand whether the taxi is wandering around after the DO instead of knowing already where to go for the next PU.   + Partition Hangzhou metropolitan area (longitude [120.0,120.5], latitude [30.15,30.4]) into 40×20 grids with equal intervals (a 1200x1200 area. * Empirical study:   + Hotspots analysis: select top 99 busiest regions.   + Between DO and PU strategy: By analyzing the average pickup numbers during one certain period of time in one certain location with respect to strategies (calculate the weight of each strategy). * Taxi pattern:   + Finding strategies: the factors are summarized in the triplet (Time, Location, Strategy), strategy is hunting/waiting (before PU) and local/distance (after DO), time and location is between DO and next PU. In this case, we have 12 (divide in 2h lags) x 100 (locations) x 2 (strategy) x 2 (PU/DO) = 4800 patterns and 4548 taxis (4548x4800 matrix).   + Strategy evaluation: adopt L1-Norm Support Vector Machines (SVM) for dimensionality reduction (finding most important features). The L1-norm regularization SVM will lead to a sparse classifier (with many feature weights equal to 0). |
| Results | * Hotspot Analysis: PU are usually in main roads while DO are wherever. Also, during night top DO are residential zones while during rush hour top PU are in the CBD. * Between DO and PU strategy: at late night and early morning (from their scattered homes to work) hunting is better, while in rush hours is better to wait. In general hunting is better but maybe is because in traffic (worst possible situation) they always look like waiting even if hunting. * Strategy evaluation: test accuracy around 85.3% (while normal SVM had 74%) and report top-10 positive and negative taxi patterns |
| Conclusion | * The paper proposes a novel method based on the triplet (time, location, strategy) to represent passenger-finding strategies. * Also, selected taxi-patterns can well interpret the empirical study results derived from raw data analysis and even reveal hidden “facts”. * Pattern prediction with 85.3 % accuracy using L1-Norm SVM |
| Way Forward (if any) | * Not mentioned * Subjective: are there strategies not used that could be even more profitable, the study only considers strategies in use. Also, there are possible nuances between hunting/waiting before PU and local/distance search after DO. |
| Key takeaways from the paper | The paper highlights the importance of understanding the strategy of a taxi driver when seeking a customer. In fact, strategies can be useful to predict future PU and DO locations. |

|  |  |
| --- | --- |
| 8 | |
| Paper Name | Inferring demand from partially observed data to address the mismatch between demand and supply of taxis in the presence of rain |
| Author | Seyyed Yousef Oleyaei-Motlagh, Adan Ernesto Vela |
| Publishing Year | 2019 |
| Case Study | New York, NY, United States |
| Dataset | * Taxi GPS dataset: TLC data from dec 2008 to Jan 2010 (400,000-500,000 daily trips) with high spatiotemporal resolution * Weather data: NOOA hourly weather information from Central Park, La Guardia and JFK ground stations * New York map |
| Aim of the paper (Abstract) | * Analyze the mismatch in supply and demand of taxis * Understand the effect on the demand of rain for taxis in New York City by looking at number of PUs, average income per drive, and empty trave time |
| Target Users | Both customers and taxi drivers can be better off by understanding how demand changes (volume, distance travelled, etc.) in bad weather situations. The problem for customers in summer rain is that they could wait for a long time or opt for the really humid metro system. |
| Main Body / Methods | * Exploratory Analysis:   + Aggregate PU/DO data to create total supply in NYC   + Use Density estimation charts to understand the distribution of shifts start and end time   + CDF of travel distance for both clear and rain settings   + Line chart explaining difference between weekday and weekend pickups   + Line charts for supply/ avg. travel speed/ avg. ride distance/ avg. taxi PUs for different time of day * Validate results:   + Nonparametric test of Mann- Whitney which is equivalent to parametric test and a nonparametric test for median using permutation analysis (instead of ANOVA since data fails normality and heteroscedasticity assumptions) |
| Results | * Average travel time is 13.5 minutes and empty time is 11.8 minutes (general setting) * Time of day is a crucial factor in predicting demand (but patterns are different in different areas of the city * The effect of rain is different on weekdays than on weekends:   + Lower increase in weekend pickup if rainy morning with less demand mismatch   + Highest number of pickups in the evening with magnitude of mismatch unclear * Shortest trips are at 6 pm * If there is a weather shock very late there will be a big mismatch * Transit ridership has peaked at evening when people come back at work and later in the mornings when people go to work |
| Conclusion | * During weekdays rain increases the taxi demand (can’t arrive to work wet) * During weekends rain decreases the need to go to museums and other recreational activities (better to stay at home) * Higher demand when it rains citywide |
| Way Forward (if any) | * Could add other weather-related factors (e.g. temperature) * Did not consider special events as a factor * Did not cross-validate analysis with external data sources (e.g. traffic) |
| Key takeaways from the paper | This paper can be a starting point for our study. It includes a substantial exploratory analysis and reaches interesting conclusions. However, we will need to dig deeper to have a prediction analysis. |

|  |  |
| --- | --- |
| 9 | |
| Paper Name | Mining the Semantics of Origin-Destination Flow using Taxi Traces |
| Author | Wangsheng Zhang, Shijian Li, Gang Pan |
| Publishing Year | 2012 |
| Case Study | Hangzhou, China |
| Dataset | Dataset provided by the Traffic Bureau of Hangzhou City, which contains 7952 taxis and covers a period of 385 days. Taxis' state is sampled in a fixed time interval of 1 minutes and an extra sampling will be performed when the taximeter turn on or off. The position was obtained by GPS equipped in a taxi, so its precision was not affected by local tower density, which limited the spatial resolution of mobile-phone data. Each state consists of following fields:   * TAXI ID: the unique ID of sampled taxi; * GPS POSITION: the longitude and latitude of that taxi at the sampling time; * SPEED: the taxi speed at the sampling time, in kilometer per hour; * ORIENTATION: the direction of that taxi at the sampling time, from 0° to 360° in clockwise with 0° indicates the north; * METER STATE: indicates whether the taxi is heavy at the sampling time, 1 means the taxi is heavy(with passenger) and 0 means the taxi is empty(without passenger); * TIME: the sampling time, with timestamp format 'YYYY-MM-DD HH:MM :SS'. |
| Aim of the paper (Abstract) | The author aims to study the GPS traces of taxis and find if there are any significant patterns under the OD flows as well as relationship with the semantics of OD flows.  Proposed an approach which offers a novel way to explore the human mobility and location characteristics; |
| Target Users |  |
| Main Body / Methods | Pattern Analysis   * The frequency f(k) of the kth most visited OD flow follows Zipf’s law,indicating most of human movements in the city occur on some major OD flows. * Graphical user interface, chart    Description automatically generated * The number of taxis’ traces shows a significant periodic pattern of 1cyc/day and 1cyc/week under the examination of power spectral density function, which reflect the short-term dynamic of city. * Human activity mainly occur on day-time and weekend for this OD flow. * Diagram    Description automatically generated with medium confidence * There are also significant difference among OD flow clusters * Chart    Description automatically generated * Semantics Mining * Define the semantics of OD flow by the semantics of its origin location and destination location such as Station to Station or Dwelling to Bar, then analyses the pattern of different semantics of OD flows. * Chart    Description automatically generated * Chart    Description automatically generated |
| Results | Use a two-layer feed-forward Neural Network with sigmoid hidden and output neurons to classify the semantics of OD flows, the prediction result shows below.  Table  Description automatically generated  The feature vectors can be define as :  Text, letter  Description automatically generated |
| Conclusion | In this paper, the author estimates the origin-destination(OD) flows from taxis' traces and find that they have significant periodic patterns which closely related with their semantics. Then they mined the semantics of OD flows based on those patterns and the experiment result achieves a recognition accuracy of 83.7%. |
| Way Forward (if any) | Future work includes analyzing the semantics change of OD flow to discover urban events, comparing OD flow's pattern under different conditions such as urban-size or develop-level and detecting communities in city via the semantics of OD flows among them. |
| Key Takeaways | * The semantic information combined with OD flow can help to better understand the pattern of taxi trajectory. * The number of taxi tracks is distributed regularly, and weekly repeated patterns can help us better mine the semantic information of OD flows. |

|  |  |
| --- | --- |
| 10 | |
| Paper Name | Time-Location-Relationship Combined Service Recommendation Based on Taxi Trajectory Data |
| Author | Xiangjie Kong, Feng Xia, JinZhong Wang, Azizur Rahim |
| Publishing Year | 2017 |
| Case Study | Beijing, China |
| Dataset | * The GPS traces dataset is generated from 12,000 taxis running in Beijing, the raw dataset collected by GPS devices contain hundreds of millions of records * To improve prediction accuracy, we clean repeated data and invalid data which the status value is 0 as shown in Table below. * In addition, we remove the logs with the status value of 2, 3 and 4, which represent the non-service taxis. The percentage of these error logs is about 0.85%   The dataset description is provided below:  Table  Description automatically generated |
| Aim of the paper (Abstract) | * Recently, urban traffic management has encountered a paradoxical situation which is the empty carrying phenomenon for taxi drivers and the difficulty of taking a taxi for passengers. * In this paper, through analyzing the quantitative relationship between passengers’ getting on and off taxis, we propose a Time-Location-Relationship combined taxi service recommendation model (TLR) to improve taxi drivers’ prof- its, uncover the knowledge of human mobility patterns, and enhance passengers’ travel experience. * Finally, we compare our proposed model with Auto Regressive Integrated Moving Average model (ARIMA), Back-Propagation Neural Network model (BPNN), Support Vector Machine model (SVM), and Gradient Boost Decision Tree model (GBDT) by using the real taxi GPS data in Beijing. The experimental results show that our optimizing taxi service recommendation can predict more accurately than others |
| Target Users | For taxi drivers, the model can help taxi drivers find the next passenger more efficiently while increasing hourly revenue.  For passengers, it can effectively shorten the average waiting time and improve the ride experience. |
| Main Body / Methods | Predicting Passenger Volume   * Partition the specific region into 6\*6 small square ones and evaluate prediction accuracy. * Then TLR model is used to predict the amount of pick- up passengers in every sub-region. * Achieve average prediction accuracy is 90.9% on weekdays and 80.4% on weekends respectively,which illustrates that the model is efficient and feasible to predict the passenger volume.   Recommending Top-N Areas   * Through utilizing the TLR model to provide the the most potential regions to drivers who drop off customers for cruising.   Performance Evaluation   * Compare the performance of TLR model with ARIMA model, BPNN model, SVM model, and GBDT model with the evaluation metric of root mean square error(RMSE) and mean absolute error(MAE), the result shows that TLR model performs best. |
| Results | * The authors propose some important human mobility patterns of functional regions through analyzing the quantitative relationship between passengers’ getting on and off taxis in every period. * The authors present TLR model, which can identify three-dimensional properties of city dynamics to predict the distribution of passengers for different social functional regions. * The authors recommend Top-N areas to drivers based on the prediction outcomes, mean trip distance, and average trip time. Then they can decide where to pick up passengers to maximize their profits. The results achieve prediction accuracies of 90.9% on weekdays and 80.4% on week- ends respectively. * The authors evaluate and compare the performance of our pro- posed model, ARIMA model, BPNN model, SVM model, and GBDT model by utilizing the following metric- s including Correlation Coefficient (CC), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Normalized Mean Absolute Error (NMAE), which determines the effectiveness and stability of our proposed model. |
| Conclusion | * In this paper, we have proposed a taxi service recommendation model named TLR by analyzing the quantitative relationship between passengers’ getting on and off taxis in different functional regions during each period. * Furthermore, we have conducted extensive simulations on TLR model and compared its performance against ARIMA model, BPNN model, SVM model, and GBDT model. The results have shown that TLR outperforms other four methods with high- er prediction accuracies of 90.9% on weekdays and 80.4% on weekends respectively. Additionally, TLR model has the lowest prediction error rate with the NAME of 15.8% on weekdays and 34.4% on weekends. |
| Way Forward (if any) | * In the future work, we will consider different social properties and multi-source datasets to improve our prediction accuracy. * We also plan to quantitatively evaluate our proposed model using bus drivers’ income data. Besides, we will focus on supply-demand matching and recommendation between passengers and taxis, which makes passengers find vacant taxis in less time. |
| Key Takeaways | * Modern cities are made up of diverse functional areas, such as commercial areas, residential areas and entertainment areas, these areas are mostly shaped by people’s actual needs for social activities. Thus, it is necessary to illustrated how to improve drivers’ profits by leveraging functional regions’ characteristics. * The evolution law of passengers’ going up and down in social functional areas is relatively stable and exclusive compared to individual mobility, and can be used as a service recommendation for the taxi drivers.   Eg. People often go to diplomacy areas for handling affairs in the morning and then prefer to have lunch rather than leave immediately, which contributes to lots of cafe and tea restaurants that serve fairly decent food. |

|  |  |
| --- | --- |
| 11 | |
| Paper Name | Understanding Taxi Service Strategies rom Taxi GPS Traces |
| Author | Daqing Zhang, Bin Li, Chao Chen, Gang Pan |
| Publishing Year | 2014 |
| Case Study | Hangzhou, China |
| Dataset | 7600 taxis served in Hangzhou, China for one year(April 2009~March 2010), including real-time taxi information such as longitude/latitude, time-stamp, passenger status(‘occupied’ or ‘vacant’), the driving speed and orientation. |
| Aim of the paper (Abstract) | Mining GPS traces to understand the taxi service strategies for taxi drivers from three main perspectives:(passengers-searching strategies, passenger-delivery strategies and service-region preference), then build up a prediction model by taxi service strategies represented with a feature matrix and evaluate the correlation between service strategies an d revenue. |
| Target Users | Understanding taxi service strategies can help taxi drivers improve their revenue, help passengers spend less time waiting for a taxi, and help city planners allocate taxi resources more rationally. |
| Main Body / Methods | * To understand the behaviors of a taxi drivers, authors first propose to separate the taxi GPS traces of each pair of shared taxi drivers based on the fact that each taxi shift handover twice a day, and the rotation occurs at a fixed time period and location. * Then the author evaluates the average performance of taxi drivers at different times of the day, such as the distribution of average hourly benefits of taxis at different time periods and the geographical distribution of population density at pick-up and drop-off locations. * After that, the authors build a model to predict the hourly earnings of drivers under different service strategies by using a feature matrix, based on analyzing the correlation between each service strategy and the average revenue |
| Results | Built up a SVM prediction model and predict the revenue of taxi drivers based on their strategies and achieve a prediction residual as less as 2.35RMB per hour. |
| Conclusion | * Generally speaking, it is found that hunting is usually more efficient than waiting in order to find passengers locally. * Going distant becomes a preferable service strategy when a taxi drops-off passengers in the suburb area, where taxi usually spend less time on average finding the next passengers by moving from non-hot areas to hot ones. * The correlation between average passenger delivery speed and the revenue shows that when the traffic becomes congested at busy time slots, choosing the light-traffic route increases the driver’s revenue. * In the end of the paper, authors obtain a residual of less than 2.35 RMB/hour, suggesting that the extracted taxi service strategies with their proposed approach well characterize the driving behavior and performance of taxi drivers. |
| Way Forward (if any) | * The author plans to broaden and deepen his work in two directions: * Conduct further research in characterizing subtle features of the taxi service behaviors and strategies, and to understand the human decision-making process. * Attempt to explore the appropriate ways which provide more concrete instructions to taxi drivers according to the taxi service. |
| Key Takeaways | * Most taxis have shared use, meaning that a car may be used by more than one driver in a day, and different drivers have different service strategies, so for this case, the driving data of each driver needs to be separated. * Most drivers vary widely at different times of the day, so it is more reasonable to analyze the performance of each driver at different times, and the data will be smoother. * The impact of a driver service strategy on profitability can be measured by correlation analysis, and a positive and larger correlation indicates that the strategy will bring more benefits to drivers, and vice versa. |

|  |  |
| --- | --- |
| 12 | |
| Paper Name | Uncovering urban human mobility from large scale taxi GPS data |
| Author | Jinjun Tang, Fang Liu, Yinhai Wang, Hua Wang |
| Publishing Year | 2015 |
| Case Study | Harbin, China |
| Dataset | **Description**:The taxi GPS data is collected from about 1100 drivers in Harbin city, which locates in the northeast of China.  **Length of period:** start from July to December in 2012, the recording rate is 30 s, and total samples come to 2880 a day.  **Feature columns:** Taxi ID, Time, Latitude, Longitude, Speed, Orientation, Status. |
| Aim of the paper (Abstract) | In this paper, we use taxi GPS data collected from more than 1100 drivers in Harbin city to characterize people travel movement. Which include:   * The distribution patterns of origins and destinations on weekday and weekend. * Using travel distance, time and speed to explore human mobility by extracting taxi trips from GPS trace data. * Verify the effectiveness of entropy-maximizing method for modeling trip distribution. |
| Main Body / Methods | * We firstly divide the city area into 400 different transportation districts and analyze the origin and destination’s taxi demand distribution in urban area on **weekday and weekend.** * Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is used to cluster pick-up and drop-off locations. * Further-more, four spatial interaction models are calibrated and compared based on trajectories in shopping center of Harbin city to study the pick-up location searching behavior(**Huff model**). * By extracting taxi trips from GPS data, travel distance, time and average speed in occupied and non-occupied status are then used to investigate human mobility. * Finally, we use observed OD matrix of center area in Harbin city to model the traffic distribution patterns based on entropy-maximizing method, and the estimation performance verify its effectiveness in case study. |
| Results | * A classical Huff model is used to analyze drivers’ choice behavior, The results show that the classic Huff model has the best fitting performance with parameters α = 1.0063 and β = −0.2812. * The probability distribution of frequency of travel distance can be fitted with two function:the first part is a power-law function and the second part is truncated power-law function, the same goes for travel time distribution but with different parameters. * The Traffic distribution is fitted by the entropy-maximizing model and get a absolute mean error of 0.0407. |
| Conclusion | * The DBSCAN algorithm is used to cluster pick-up and drop-off locations, and two key parameters (MinPts and Eps) in the algorithm are optimized. * The classical Huff model has the best modeling accuracy among those four spatial interaction models. * The distribution of taxi trips in occupied status include two patterns: ascending part and descending part. The distribution in ascending part can be well fitted by power law function, and curve in descending part is followed truncated power law function. * As to distribution of trips in non-occupied status, there only exists a monotonically pattern, which can be fitted by truncated power law function. * Optimizing the parameters in entropy-maximizing model based on actual OD matrix and evaluate its estimation accuracy for traffic distribution. |
| Way Forward (if any) |  |
| Key Takeaways | Comparing with grid based method, using DBSCAN algorithm to cluster PU/DO locations has these benefit:   * DBSCAN is a spatial density based method, and it can classify the locations in a cluster with high density, also the specific locations in each clusters can be found in road network. * DBSCAN can filter out the interfering noise. In China, the passengers frequently call for vacant taxis in the middle link of road, as these pick-up locations appear randomly. DBSCAN can realize this function through selecting proper parameters. * The classical Huff model has the best modeling accuracy among those four spatial interaction models. |

|  |  |
| --- | --- |
| 13 | |
| Paper Name | Inferring Passenger Denial Behavior of Taxi Drivers from Large-Scale Taxi Traces |
| Author | Sihai Zhang\*, Zhiyang Wang |
| Publishing Year | 2016 |
| Case Study | Beijing, China |
| Dataset | **Description**: 12,657 taxis equipped with GPS devices during a period of two  months. The sampling interval ranges from 10 seconds to 60 seconds, depending on the settings of different taxis. There are totally around 2 billion rows of records.  **Length of period:** from November 1st to December 31st, 2012  **Feature columns:** Taxi ID, Time Stamp, Operation status and trigger, Latitude, Longitude, Speed, Orientation, GPS Status. |
| Aim of the paper (Abstract) | * finding out the actual income of each taxi driver * grouping taxi drivers into different income level with proposed diversity concepts * understanding the pick-up and drop-off patterns for each group.We filter out four groups of taxi drivers according to their pick-up and drop-off diversity, and each group comprising drivers with High income(525 drivers),Medium high(542 drivers),medium low(510 drivers) and low(516 drivers), respectively. * We demonstrate that the four groups exhibit different pick-up and drop-off patterns * and that the high income drivers exhibit passenger denial behaviors. |
| Main Body / Methods | * Based on the complete trace information for taxis supplied by one Chinese Investigation Agency, we study the income differentiation, pick-up diversity, drop-off diversity and grid diversity in all taxi drivers. Our works are performed at both the individual and group levels. * To ensure better statistics, the top 3,590 taxis with the largest number of served passengers are chosen as our data sample, each having more than 932 single trips and millions of GPS report records. We propose a bottom-up approach to investigate individual taxi driver’s pick-up behavior. * Painting single driver’s trajectory with longitude and latitude. * Bottom-up dividing drivers into groups according to income level. * Understanding pick-up and drop-off location patterns of each group, drawing PDF of daily net and gross incomes. * Finding out difference among groups on passenger denial. |
| Results | * The so-called mobility intelligence do not necessarily increase the income of taxi drivers, unless they choose the proper waiting areas. Here, high income and low income taxi drivers are two opposite examples. * High income taxi driver are exposed to deny passengers after knowing their destinations and the estimated denial rate of high income taxis is 8.52%. |
| Conclusion | * Drivers in different income levels exhibit different patterns on pick-up and drop-off locations. High-income drivers are more likely to deny passengers according to their preference on locations. |
| Way Forward (if any) | We can explore the dark side investigation on human dynamics. |
| Key Takeaways | Look into the problem that will preferential strategy increase drivers’ income, the factors affecting high-income drives, and factors about why drivers deny passengers based on proportional distribution and visualization. Find drivers’ pick-up and drop-off patterns by grouping locations into grids to streamline the geographic characteristics. |

|  |  |
| --- | --- |
| 14 | |
| Paper Name | Full Cycle Campus Life of College Students: A Big Data Case in China |
| Author | Benyou Wang, Kaihe Deng, Weiwei Wei, Sihai Zhang, Wuyang Zhou, and Shui Yu |
| Publishing Year | 2018 |
| Case Study | Hefei, China |
| Dataset | **Description**: The data set used in this paper is collected by the campus smart card(CSC) system. CSC takes the intelligent IC card as the medium for identity authentication, information storage, stored-valued consumption, and information transmission, with the support of the computer database technology and network communication  technology. And the system records the time, the student id and other information needed.  **Length of dataset:** The data set includes 4,600 students, 1,832 female and 2,697 male, enrolled in 2012 and graduated in 2016. But there are 71 students who dropout or delay their graduation for different reasons, who are excluded in this paper. Then, the total number of students investigated in this paper is 4,529.  **Feature columns:** Card id, the consumption time stamp, the consumption amount(paying or recharging), the consumption account and so on. |
| Aim of the paper (Abstract) | * find the characteristic of college students’ life with behavior analysis using big data method. We want to find the trend of behavior changing. We want to grasp of the overall situation of students’ life and use the result to promote our further work. * analyze the students’ behavior by the multi-dimensional view and prove some conclusion which used to be made by education experience. |
| Main Body / Methods | * As described above, there are totally 37 POS accounts, from which we can understand the major consumptions of students in campus. Firstly, the plot of food consumption in canteens, presents the average consumption amount of each   time and daily consumption amount of each student, together with the daily sum of consumption amounts of all students, which also tells us several findings.   * By differentiating the two gender groups, it presents the food behavior of college students, showing the difference between boys and girls, breakfast, lunch and dinner. Several findings are discussed here. (1) Difference between meal kinds. The students’ breakfast frequency is much lower than lunch and dinner. (2) Gender difference. |
| Results | * We think girls may be better at purchasing items in different ways outside the campus to find the most satisfying goods, while boys are more likely to choose the most convenient way to purchase items. * For the second finding, we can infer that male students like surfing Internet more than female students, the same as the common sense in education field. * Girls are more hardworking so that they spend more time on reading books and there might be more literature lovers among girls than boys. Using the detailed books categories in the borrowing records, we may verify this in our further works. * As to the vacation effect, we can infer that almost 60% students leave the campus, for returning home, traveling outside or for other things, which leads to the decrease in consumption. As for price increasing, the yearly increase of China’s food and commodity price is about 3%, so it shows that students tend to choose more expensive consumption in campus, which shows the life quality of students in campus is improving. |
| Conclusion | * The consumption in campus has a clear time-dependent law. Consumption reduced a lot even to 1/6 of an ordinary day in some special days. * There is a changing trend of students’ behavior in the learning career and the behavior in campus shows certain difference between two genders. The consumption of students shows an increase in many aspect and students of different genders show different living habits. And from the library data we find that the life of students’ last year in campus may be much more busy than the first 3 year. |
| Way Forward (if any) | Through the data analysis and forecast we can make reasonable arrangements, which can not only guarantee the normal operation, but also make rational deployment of resources, improve efficiency, reduce waste. |
| Key Takeaways | Contrastive analysis between different school years and genders can be of great use to consumption patterns. Time series analysis on the term scale also tells a good story on special cases like weekends and holidays. |

|  |  |
| --- | --- |
| 15 | |
| Paper Name | Taxi Driving Behavior Analysis in Latent Vehicle-to-Vehicle Networks: A Social Influence Perspective |
| Author | Tong Xu, Hengshu Zhu, Xiangyu Zhao, Qi Liu1 Hao Zhong, Enhong Chen, Hui Xiong |
| Publishing Year | 2016 |
| Case Study | New York City |
| Dataset | **Description**: The data set e taxi driving transactions in New York City during the whole year of 2013, which is provided by NYC Taxi and Limousine Commission (NYC TLC).  **Length of dataset:** This is a largescale data set that totally consists of more than 169 million transaction records of 43,191 drivers in 14,144 cabs. For each transaction, we have the spatial and temporal information for both pick-up and drop-off, as well as fares including tip and toll.  **Feature columns:** Term, Transactions, Speed, Income. |
| Aim of the paper (Abstract) | * As we propose the idea of driving behavior propagation within cab drivers, in this section, we will intuitively discuss the effects of social propagation with related statistical analysis to further support our motivation. * Intuitively, if we treat taxi drivers as “social agent” in the mobile social networks, and simulate how the “social propagation” scheme functions to interpret their future behaviors, taxi route will be more predictable, and further social-oriented taxi services, e.g., social-based “tutor” or pattern recommendations could be effectively conducted. |
| * Main Body / Methods | * Further illustrates the motivation with some intuitive statistics. * Propose the novel framework for driving behavior analysis with integrating social factors. * Measure correlations of driving skills, here we choose three evaluation metrics, i.e., the transaction amount, the average driving speed and the total income, to study on this issue. For each evaluation metric, the top 10 drivers are ranked, compared with other 10 randomly selected drivers. * Two-stage loss function optimization:   + Training Stage. Given a group of taxi drivers U = {ui} as well as their pattern vectors s t i during the time period t = 1, 2, ..., T, in the training stage, we aim at inferring the latent social connections {wij} within drivers, which achieve the best explanation for the ranking of driving behavior vector fluctuation ∆sT+1 i .   + After obtaining the latent connections {wij} in the training stage, in the test stage, given the taxi drivers group U = {ui} with their pattern vectors s t i during the certain p-time lag as t = T − p + 1, ..., T − 1, T, we aim at predicting the driving behavior vector fluctuation ∆sT+1 i with accurate sign and ranking. * To generalize the driving patterns, we first clustered all the pick-up and drop-off locations in the historical transaction records. Specially, we conducted a bottom-up hierarchial with minimum variance criterion until only 30 clusters were left. * To predict the pattern change, we indeed have two tasks, i.e., the binary classification to distinguish the sign (positive / negative) of pattern increment, and then ranking the patterns with respect to their increments. |
| Results | * The overall prediction performance of our approach SPC (Social-aware Pattern-Change prediction) and other baselines. To be specific, the top 300 patterns were studied and the time lag was set as 5 months. we realize that behavior patterns of cab drivers could be largely random, as all the performance are relatively poor. However, we can find that except for the comparison with VAR on ranking problem, our approach outperforms the other baselines with dramatic margin, even 20 times better in some experiments. * For VAR model, it performs truly great in ranking task, but terribly fails for binary classification. With deep looking of the output of VAR, we realize that usually VAR predicts the proportion as 0 or negative, not only for those patterns that the drivers never try, but also for those drivers tried for once but never reappear. |
| Conclusion | * The heuristic methods might not be appropriate to estimate driving patterns of taxi drivers if without considering additional factors, like financial benefits or running speed. This phenomenon might further explain why our model could outperforms the baselines, as we do not “teach the model” how to predict the change, but intuitively “simulate the social propagation scheme”, which is finally proved as effective. Clearly, except for those intellectual services, taxi drivers themselves could be the “best learner”. |
| Way Forward (if any) | Though social factors could better explain the pattern fluctuation at most for around 80%, there are still some other key factors, e.g., financial profit or traffic environment. In the future, we would like to investigate these factors with more comprehensive prediction framework. Also, social-oriented taxi services, e.g., social-based “tutor” or pattern recommendations will be considered. Finally, we will discover whether similar solutions could be used for other service-oriented professions. |
| Key Takeaways | Mobile Data Mining, Binary Classification, Optimization |

|  |  |
| --- | --- |
| 16 | |
| Paper Name | Variable Selection for Personalized Driving Behavior Modeling and Application to Autonomous Driving |
| Author | Jude Chibuike NWADIUTO |
| Publishing Year | 2021 |
| Case Study | New York City |
| Dataset | **Description**: The data set e taxi driving transactions in New York City during the whole year of 2013, which is provided by NYC Taxi and Limousine Commission (NYC TLC).  **Length of dataset:** This is a largescale data set that totally consists of more than 169 million transaction records of 43,191 drivers in 14,144 cabs. For each transaction, we have the spatial and temporal information for both pick-up and drop-off, as well as fares including tip and toll.  **Feature columns:** Term, Transactions, Speed, Income. |
| Aim of the paper (Abstract) | * To solve the existing issues in the hybrid dynamical systems (HDS) modeling for understanding the dynamical characteristics of the human driving behavior specifically answering the questions:   + How to decide the optimal number of modes/behaviors in a HDS when modeling the human driving behavior?   + Explain the switch between modes in a HDS when modeling the human driving behavior? |
| Main Body / Methods | * Piecewise Autoregressive Exogenous (PWARX) model * Weighted k-means clustering algorithm and variable selection * Logistic regression and variable selection * Support Vector Machine SVM and variable selection * Model evaluation |
| Results | * A three-mode model was identified as the optimal number of modes amongst the drivers for the car-following task in the downtown area. * The car-following driving task in the downtown area can be said to consist of three modes or sub-models, (i.e., the dangerous region, safe region, and cruising region. However, the separating hyper-planes (decision boundaries) are different from one driver to anothe. * Each driver has his own unique model structure (i.e., unique decision making) for the car-following task, and thus this uniqueness can be leveraged directly to design automated systems 45 peculiar to each driver. |
| Conclusion | * By leveraging the idea of consistent variable selection, an optimal model can be realized. The PWARX model is well suited to describe both the decision-making and motion control facets in analyzing human driving behavior. * The proposed model shows significantly better prediction performance and is able to mimic the real-road driving behavior better when simulated. |
| Way Forward (if any) | In the design of an advanced driver assistance system (ADAS) for autonomous driving tasks, driving situations can be identified by using the values of specific explanatory variables or by looking at the values of decision variables by using the 82 work presented here. |
| Key Takeaways | hybrid dynamical systems (HDS) modeling, Variable Selection |

1. https://github.com/fivethirtyeight/uber-tlc-foil-response [↑](#footnote-ref-1)