Analyzing the Gap Between Ride-hailing Location and Pick-up Location with Geographical Contexts

Yunlei Liang
Geospatial Data Science Lab,
University of Wisconsin, Madison
yunlei.liang@wisc.edu

Song Gao
Geospatial Data Science Lab,
University of Wisconsin, Madison
song.gao@wisc.edu

Mingxiao Li
Institute of Geographic Sciences and
Natural Resources Research, Chinese
Academy of Sciences
limx@lreis.ac.cn

Yuhao Kang
Geospatial Data Science Lab,
University of Wisconsin, Madison
yuhao.kang@wisc.edu

Jinmeng Rao
Geospatial Data Science Lab,
Universityo f Wisconsin, Madison
jinmeng.rao@wisc.edu

ABSTRACT

Location uncertainty plays a key role in location-based services and applications. This research mainly focuses on the problem of location uncertainty when users are using the ride-hailing applications from the perspective of geographical contexts. The distance gap between the ride-hailing identified search location and the actual pick-up location was calculated and used as the measurement of location uncertainty. It may come from GPS noise or the gap between the current location of a passenger (e.g., inside a building or a university campus) and the actual pick-up location along the streets for a ride. For the geographical contexts, this study considers factors including population density, road density, various kinds of Points Of Interests (POIs), and also the frequency of ride-hailing requests given a region. By using regression analysis techniques to find the relation between the geographical contexts and the distance gap, this study identifies potential factors (e.g., enterprise buildings, dense roads, and highly populated areas) that affect the location uncertainty the most.

CCS CONCEPTS

• **Information systems → Geographic information systems**; **Location based services**.

KEYWORDS

location accuracy, ride-hailing, regression analysis

ACM Reference Format:
Yunlei Liang, Song Gao, Mingxiao Li, Yuhao Kang, and Jinmeng Rao. 2019. Analyzing the Gap Between Ride-hailing Location and Pick-up Location with Geographical Contexts. In *1st ACM SIGSPATIAL International Workshop on Ride-hailing Algorithms, Applications, and Systems (RAAS'19), November 5, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3357140.3365493

RAAS'19, November 5, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6969-5/19/11…$15.00
https://doi.org/10.1145/3357140.3365493

1 INTRODUCTION

The study of positioning accuracy and location uncertainty is a crucial yet challenging problem for ride-hailing applications (e.g., Uber, Lyft, and DiDi) and many other location-based services (LBS) [6, 7]. The positioning component determines the current location of a user. For outdoor LBS applications, assisted-GPS is often used for positioning on mobile devices with multiple sensors (e.g., GPS, cellular receiver, and WiFi). Zandbergen (2009) reported that assisted-GPS positioning obtains an average median error of 8m outdoors, while WiFi positioning gets 74m of that and cellular positioning is the least accurate with about 600m median error [20]. For indoors, GPS positioning doesn't work well because of the signal obstructions and attenuation. Thus, Wireless Area Local Networks (WLAN), Radio-Frequency Identification systems (RFID), and low-energy Bluetooth sensor networks are often used as indoor positioning infrastructures [5, 13]. Besides, researchers have made great efforts to improve the GPS positioning accuracy and trip distance estimations. Examples include lessening signal multipath propagation in urban canyons and tracing multiple candidate positions [14], using an improved positioning architecture without hardware-level modification [19], or taking the patterns of GPS noise levels [15] and sampling rates [11] into consideration .

The locations of ride-hailing users are important but often inaccurate especially when users are inside buildings, underground parking lots, elevators or even on the street that is surrounded by skyscrapers [9]. As a result, the ride-hailing applications sometimes cannot identify the correct location of the user or the location is not accessible by vehicles due to transportation regulation (e.g., inside an enterprise park or a university campus). In such cases, most of the users would type in a pick-up location around their actual locations in order to get the ride or the ride-hailing application may recommend a more accessible pick-up location instead. The gap between the users' actual location and the location identified by the ride-hailing applications might contains useful information reflecting the reasons why such inaccuracy appears. Therefore, by computing the distance gap between the location identified by the rider-hailing application and the user's actual pick-up location and analyzing potential geographical factors that are related to this gap, this research provides insights into further improvement of the location accuracy in the ride-hailing applications.

2 DATA AND METHODS

2.1 Study area and Data

The study area is in Chengdu, a major city in the southwestern part of China. The dataset is from the GAIA Initiative provided by DiDi Chuxing, a well-known ride-hailing transportation company in China [1]. The data includes large-scale location search requests in May, 2018 and has the information of the identified ride-hailing location as well as the pick-up location. There are over 650,000 records in total.

Given the research interest, this study only looks at the requests that are for trip origins and there are 1732 records in total used in the analysis.

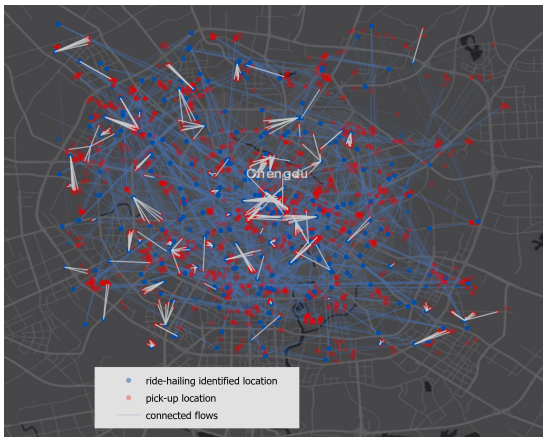

Figure 1: The gaps between ride-hailing identified locations and pick-up locations. Lines in white represent the distance gaps within 1.5 km.

2.2 Data processing

The location data from DiDi is first filtered to obtain all ride-hailing requests for the origins. The data is transformed from the original GCJ-02 coordinate system to the WGS84 system for associating with other geospatial datasets such as the OpenStreetMap data. Then the great-circle distances are computed for each pair of locations (e.g., the search location identified by the ride-hailing application and the pick-up location inputed by the user or recommended by the App). Figure 1 shows the connection of each pair of location on the map.

For the geographical contextual features (as shown in Figure 2), we consider factors including population density, road density, POIs, and the frequency of ride-hailing requests given a region. Each of them is considered as a map layer to be spatially overlaid on the top of the Didi location data. Then, all the geographical features are aggregated to the 1km x1km grid level respectively and the sum of each feature is calculated to be used as the derived feature value of each grid. For the POI features, we get 17 categories of POIs from the *Baidu Maps*, including beauty, business, car service, education, enterprise, entertainment, fitness gyms, food, government, hospital, hotel, house, life services, media, shopping, tourism and transportation. Each POI category is considered as a single feature layer. After aggregating each feature to the grids, the features are joined to each pick-up location so that further analysis will be between the distance gap and the geographical features at the location point level.

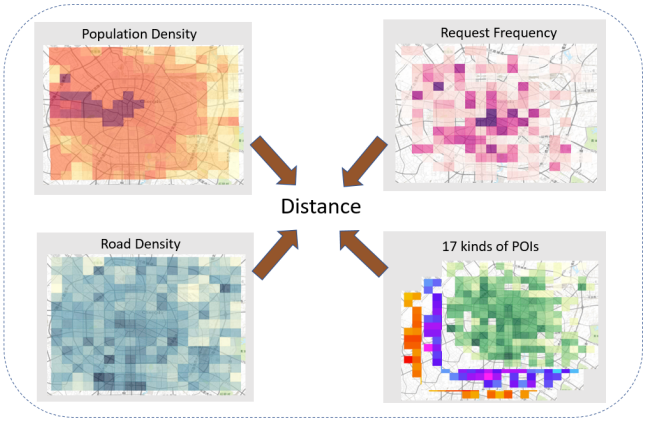

Figure 2: The analytical framework of the distance gaps between pickup locations and search locations using geographical contexts.

2.3 Regression Analysis

Two types of regression analysis methods are used in this study, namely the multi-linear regression (MLR) and the least absolute shrinkage and selection operator (Lasso). Those two methods are aimed to examine the potential relationships between variables at the initial stage. The relationship between the dependent variable and the independent variables may be more complex and more complicated methods such as non-linear models will also be applied to help further analyze the data in the future,.

The MLR is one of the most widely used regression techniques for modeling and predicting. It is able to predict the linear outcome of a dependent variable based on multiple independent variables [3]. The goal of MLR is to minimize the sum of squares of errors (SSE). In this study, the distance between each pick-up location and the search location identified by the ride-hailing applications is used as the dependent variable. The aforementioned four types of geographic features (in total 20 features) are used as the independent variables. The formula is listed below:

$$Distance = \beta_0 + \beta_1 * Pop + \beta_2 * Road + \beta_3 * Request + \sum_{i=4}^{20} \beta_i * POI_i \quad (1)$$

where β_0 is the intercept term and the remaining β_i are coefficients for each independent variable calculated from the regression model.

The Lasso function applies a $L1$ penalty term to control the coefficient of each independent variable. Given observations $x_i \in R^p$ and the responses $y_i \in R, i = 1, ..., N$, the objective function of Lasso is to minimize the following formula for each pair of (β_0, β) in R^{p+1}:

$$1/2N \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda ||\beta||_1 \quad (2)$$

where $\lambda >= 0$ is a complexity parameter [2, 4]. Here β_0 is also the intercept while β is a coefficient parameter for the independent

variable. With a higher coefficient, the penalty will be bigger and the final magnitude of coefficients will be smaller [18]. By shrinking the parameters, the Lasso regression is able to control multicollinearity. It tends to pick one from a few very correlated predictors and set the coefficients of the others to zero [2, 4]. Therefore, the Lasso regression is very helpful when there are many intercorrelated features and feature selection is necessary [18] .

In addition, the importance ranking of each feature in explaining the variance of the distance gap is generated using the Random Forest method. The reason of not directly using the coefficient values from MLR for ranking is that MLR cannot prevent the multicollinearity among independent variables and makes the coefficients may not be very reliable. For Lasso, the parameters are standardized in the model training process but are not returned, so the coefficients cannot be compared directly [2]. The variable importance refers to how much the error would drop for a variable at each split point in decision-tree methods [8]. We chose the decrease in mean-squared-error (MSE) for a variable at each split point and then averaged over all trees [17]. In the process of building a regression tree, the independent variables space is divided into a few subregions and the goal is to find region divisions $R_1, R_2, ..., R_J$ that minimize the SSE:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \quad (3)$$

where $\hat{y}_{R_j}$ is the mean response for the training observations within the j^{th} region [17].

3 RESULTS

3.1 Location Data Analysis

In the training data which has over 650,000 records, the search requests for trip origins only account for around 0.2% of the total data. One possibility is that the ride-hailing applications may automatically locate the users accurately in most cases so there is no need to search for a pick-up location but only for a destination location. However, in the 1,732 requests for origins, the distance gaps have a very large range and show a relatively high average value around 4km, which is far beyond the normal GPS error. It indicates that most of these ride-hailing search locations were not identified correctly or not accessible. The histogram and cumulative distribution function of the distance gaps are shown in Figure 3.

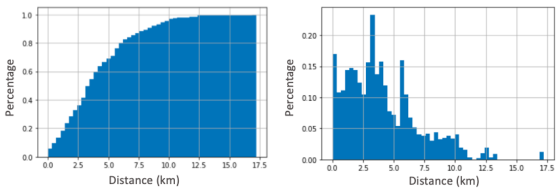

Figure 3: The cumulative distribution function and the histrogram of the distance for all origin requests.

3.2 Distance Gap Analysis

Therefore, our analysis only focuses on the data which has a distance gap that is relatively smaller and can be considered as location mismatch related to the particular location environment rather than some uncontrolled reasons. A few distance thresholds are selected to compare the results. The minimum distance threshold is set to be 500 meters, which has been used as a reasonably accepted walking distance for most of the people [10, 12, 16]. We then increase the threshold incrementally to see how the results may change.

To obtain a reliable result, this study uses a 5-fold cross validation. The data is split into 5 folds. Each time, 4 folds of data is used as the training data and the rest 1 fold is used as the testing data. The R^2 is used as the accuracy measurement. Table 1 shows the average R^2 from the cross validation of MLR and Lasso under different distance thresholds. Generally, the prediction accuracy drops when we increase the distance threshold. When we only analyze the data with distance gap smaller than 500 meters, we obtain a relatively high R^2 for the training data (MLR: 0.43 and Lasso: 0.415). The results from MLR and Lasso are different mainly because the principles of the algorithms are different. The Lasso algorithm is able to control multicollinearity and only picks part of the independent variables. It may lose some information and make the result from Lasso a little lower than that from the MLR. However, the result still shows that the proposed geographic factors could help explain the variability of the distance gaps to certain degree. For the testing data, the results are not very stable. Therefore, the average R^2 is low. The results might be caused by the small dataset size (around 200 points). Another potential reason is that the dimensions of features are relatively large compared with the small sample size. Therefore, the two regression methods may not perform well in this case.

Table 1: The regression results of MLR and Lasso under different distance thresholds (average R^2).

	500 m	800 m	1000 m	1200 m	1500 m
MLR training	0.430	0.432	0.359	0.296	0.241
Lasso training	0.415	0.361	0.341	0.294	0.224
MLR testing	0.266	0.245	0.183	0.108	0.107
Lasso testing	0.248	0.202	0.114	0.07	0.086

3.3 Variable Importance

The importance ranking for each variable is computed and shown in Figure 4. When the variable has a higher importance ranking, it would contribute more to the predicted variable "distance gap". In other words, a higher value of this variable will lead to a more frequent existence of location inaccuracy. The top importance feature is the number of *Enterprise* POIs. Since many companies are located in central areas of the city and often located inside the high buildings, it may frequently lead to location inaccuracy. The second and third important features are road density and population density. This corresponds to our life experience that the downtown areas always have dense road networks due to the fact that it needs to be accessed by a great number of people from other places. The *food* POIs help contribute to the fact that usually busier areas are more likely to have location inaccuracy issue. For the education variable, this can be related to some

phenomena in China. For instance, one passenger needs to walk to the main gate of a university campus to get on a ride due to the university transportation regulation. This situation will also cause the occurrence of large distance gap. The remaining variables will also contribute to the location inaccuracy since they all have positive importance values.

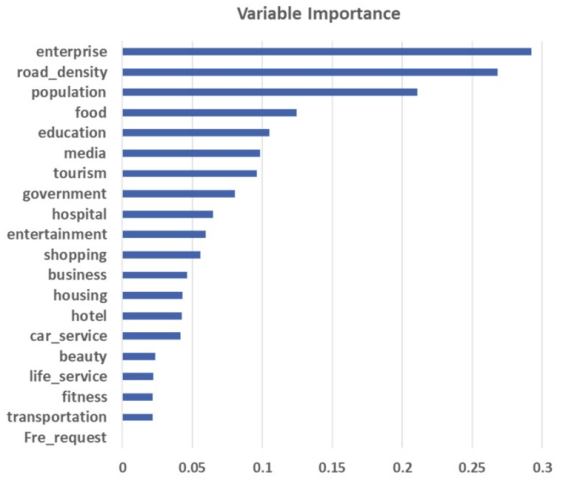

Figure 4: The variable importance ranking result.

3.4 Outlier Analysis

We also looked at the data that has very large distance gap, which can be identified as 'outliers' and tried to find whether there is any pattern leading to such occurrence. We selected the last quarter of data that has the greatest distance gaps and conducted the two regression analysis methods to them. By looking at the data distribution, the threshold for the last quarter is 5.75 km. The r^2 from MLR is 0.272 and the r^2 from Lasso Regression is 0.303. The low r^2s reflect the fact that there might be multiple reasons leading to the existence of 'outliers'. For example, it is possible that those requests that not real-time ride-hailing requests for the users. For example, they might schedule requests for future trips or they are requesting services for their friends or family who are far away from them. In addition, there might be complex reasons that cannot be included and explained in a linear relationship, which needs further investigation with additional variables or other models in future work.

4 CONCLUSION

This study analyzes potential reasons why the distance gap caused by the location inaccuracy in the ride-hailing applications exists from the perspective of geographical contexts. Factors that are taken into consideration include population density, road density, the number POIs in different categories, and the frequency of ride-hailing requests. By analyzing data that has a distance gap less then 500 meters, this study is able to find some most influential factors to the location inaccuracy such as existence of enterprises, dense road network, dense population, education area, etc. However, due to the small dataset, the result for the testing data is not very stable and leads to some uncertainty of this study. In the future, this study can be applied to larger datasets to generate more comprehensive results. Also, a few more factors such as elevation and land use types can be incorporated, which may be particularly helpful for ride-hailing applications in some mountainous areas.

ACKNOWLEDGMENTS

The ride-hailing data source is provided by Didi Chuxing GAIA Initiative. Support for this research was provided by the University of Wisconsin - Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (WARF).

REFERENCES

[1] DiDi Chuxing. 2019. GAIA Open Dataset. https://gaia.didichuxing.com [Online; accessed 13-September-2019].
[2] Lei Dong, Carlo Ratti, and Siqi Zheng. 2019. Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences* 116, 31 (2019), 15447–15452.
[3] Norman R Draper and Harry Smith. 1998. *Applied regression analysis*. Vol. 326. John Wiley & Sons.
[4] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33, 1 (2010), 1.
[5] Song Gao and Sathya Prasad. 2016. Employing spatial analysis in indoor positioning and tracking using wi-fi access points. In *Proceedings of the Eighth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*. 27–34.
[6] Haosheng Huang and Song Gao. 2018. Location-based services. *The Geographic Information Science & Technology Body of Knowledge*. (2018).
[7] Haosheng Huang, Georg Gartner, Jukka M Krisp, Martin Raubal, and Nico Van de Weghe. 2018. Location based services: ongoing evolution and research agenda. *Journal of Location Based Services* 12, 2 (2018), 63–93.
[8] Grant Izmirlian. 2004. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *ANNALS-NEW YORK ACADEMY OF SCIENCES* 1020 (2004), 154–174.
[9] Vaishnav Kameswaran, Jatin Gupta, Joyojeet Pal, Sile O'Modhrain, Tiffany C Veinot, Robin Brewer, Aakanksha Parameshwar, Jacki O'Neill, et al. 2018. 'We can go anywhere': Understanding Independence through a Case Study of Ride-hailing Use by People with Visual Impairments in metropolitan India. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 85.
[10] Thomas J Kimpel, Kenneth J Dueker, and Ahmed M El-Geneidy. 2007. Using GIS to measure the effect of overlapping service areas on passenger boardings at bus stops. *Urban and Regional Information Systems Association Journal* 19, 1 (2007).
[11] Yang Li, Dimitrios Gunopulos, Cewu Lu, and Leonidas Guibas. 2017. Urban travel time prediction using a small number of GPS floating cars. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 3.
[12] Yunlei Liang, Song Gao, Tianyu Wu, Sujing Wang, and Yuhao Wu. 2018. Optimizing Bus Stop Spacing Using the Simulated Annealing Algorithm with Spatial Interaction Coverage Model.. In *IWCTS@ SIGSPATIAL*. 53–59.
[13] Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. 2007. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, 6 (2007), 1067–1080.
[14] Shunsuke Miura, Li-Ta Hsu, Feiyu Chen, and Shunsuke Kamijo. 2015. GPS error correction with pseudorange evaluation using three-dimensional maps. *IEEE Transactions on Intelligent Transportation Systems* 16, 6 (2015), 3104–3115.
[15] James Murphy, Yuanyuan Pao, and Asif Haque. 2017. Image-based classification of GPS noise level using convolutional neural networks for accurate distance estimation. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*. 10–13.
[16] Alan T Murray and Xiaolan Wu. 2003. Accessibility tradeoffs in public transit planning. *Journal of Geographical Systems* 5, 1 (2003), 93–107.
[17] Suggests RColorBrewer and Maintainer Andy Liaw. 2018. Package 'randomForest'. (2018).
[18] SHUBHAM JAIN. 2017. A comprehensive beginners guide for Linear, Ridge and Lasso Regression in Python and R. https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/ [Online; accessed 14-September-2019].
[19] Rui Xu, Wu Chen, Ying Xu, and Shengyue Ji. 2015. A new indoor positioning system architecture using GPS signals. *Sensors* 15, 5 (2015), 10074–10087.
[20] Paul A Zandbergen. 2009. Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS* 13 (2009), 5–25.