Biometrics 000, 000–000 DOI: 000
000 0000

Unifying NYC Taxicab Records and Hotel Occupancy Data

Daniel J. Saunders

College of Computer and Information Sciences, University of Massachusetts, Amherst, Massachusetts

email: djsaunde@cs.umass.edu

and

Christian Rojas

Department of Resource Economics, University of Massachusetts, Amherst, Massachusetts

email: rojas@resecon.umass.edu

and

Debi Mohapatra

Department of Resource Economics, University of Massachusetts, Amherst, Massachusetts

email: dmohapatra@umass.edu

Summary: A method for fast distributed processing of New York City taxicab trip records in relation to a secondary dataset of hotel information is presented. An algorithm used to select an optimal distance threshold for capturing relevant trip records is developed. We show exploratory visualization of taxicab trip records in relation to hotel locations. We suggest and preliminarily investigate applications of the New York City taxicab trip records and its unification with hotel occupancy information.

Key words: Applied economics; Data analysis; Distributed computing.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

The availability of various forms of empirical commercial and industrial data is crucial for the evaluation and solution of real-world economics problems. At the scale of cities and nations, however, certain data present major data processing challenges. With the advent of massively parallel and relatively cheap distributed computing, however, methods for the processing and analysis of such datasets are not entirely out of reach.

The New York City (NYC) Taxi & Limousine Commission (TLC) Trip Record Data consists in part of yellow and green taxicab trip records spanning the years 2009-2017. Recorded attributes include pick-up and drop-off date and time of day, pick-up and drop-off geospatial or zone-coded location, trip distances, fares, and others.

Although interesting studies have been done on the taxi trip records dataset on its own, combining it with additional relevant datasets may allow for more powerful economics studies. We study a unification of the taxi data with the occupancy rates of 178 NYC hotels from the years 2013 to 2016. We are interested in trips which *begin or end within distance d* of at least one hotel in the dataset. To find the best choice of d over all hotels, we demonstrate an optimization procedure motivated from a simple distribution matching approach. Hotels with abnormal amounts of taxicab traffic may be eliminated from consideration during or after the optimization process.

We hypothesize that both hotel occupancy rates and pricing can be predicted from daily nearby taxicab pick-up and drop-off distributions. We hope that hotel occupancy and (indirectly) pricing will be a simple stochastic function of the proportion of nearby taxicab density. Indeed, our data analysis demonstrates this on a subset of hotels within a tolerable error.

2. Related Work

An analytics model is proposed in Ferreira et al. (2013) which allows users to visually query taxi trips. Standard queries about the data are supported, as well as spatio-temporal "origin-destination" queries, allwoing users to discover mobility patterns throughout the city. NYC hotspots are identified in Stoyanovich et al. (2017), in which taxi trips are represented as straight-line trajectories from pick-up to drop-off coordinates. Hotspots are used to identify lack of convenient public transportation options, and to suggest the addition of bus routes and ride-sharing options. A custom software approach is developed on top of the distributed processing system Apache Spark, allowing efficient analysis of taxi trajectory graphs as they evolve. In another work, a visual analytics method is proposed to study urban mobility patterns using graph modeling of taxicab trajectories (Huang et al., 2016). Dynamics of Shenzhen taxis is analyzed using graph analysis, and is studied at multiple scales by using a graph partitioning algorithm to produce region-level visual analytics. Users of the system can interactively explore street-level city traffic patterns.

A technique for event-guided exploration of large spatio-temporal data is introduced in Doraiswamy et al. (2014), on the basis that manual exploration of such data is time-consuming and ineffective. Computational topology is used to discover interesting events in the data, and an algorithm is developed to group and index events which can be interactively explored or queried by users. This approach is validated on NYC taxi trips as well as subway service records. An analysis of the NYC taxi trips database in combination with various relevant urban data is presented in Wu et al. (2016). Correlations between taxi and the external datasets are used to the predict traffic dynamics; namely, it it shown that points-of-interest (POIs) can predict regular traffic patterns, geo-tagged tweets can explain traffic caused by atypical events, and weather data may explain abnormal drops in traffic density. Chu et al. (2014) develop a method to discover and analyze information contained in a

massive dataset of taxi trajectories. Geospatial coordinates are replaced with traversed street names, which are subsequently modeled with topic modeling tools. The "hidden themes", or the discovered topics from the employed topic models, are used to analyze mobility patterns with a visual analytics system.

The efficient processing of spatio-temporal datasets (e.g, NYC taxi trip records) often requires specialized software and hardware. A novel indexing scheme over spatio-temporal data is developed for use with general-purpose graphics processing units (GPUs) in Doraiswamy et al. (2016). The index allows sub-second query speeds over large spatio-temporal datasets such as the NYC taxi data, and a massive number of tweets collected from Twitter.

3. Methods

3.1 *Data processing*

The dataset of hotel occupancy rates contains information from January 2013 until present day. The dataset of NYC hotel taxi records contains data from January 2009 until present day; however, geospatial coordinates of pick-up and drop-off locations are only recorded from January 2009 until June 2016. Given these constraints, our experiments consider only those records which lie in the date range of January 2013 until June 2016, approximately 3.5 years of joint data. Our analyses are concerned only with the yellow and green taxi services in NYC, and we discard all data concerning "for-hire vehicle" services (e.g., Uber, Lyft, and others).

We use a supercomputing cluster in which users are provisioned with a maximum of 40 CPU compute nodes. On each node, a Python process is dispatched to *pre-process* a single month's worth of taxicab trip data with respect to the set of NYC hotels of interest. In particular, upon specifying a *distance criterion* d, each process will extract trips from its month worth of data which begin or end with d feet of any of the hotels under consideration.

Geospatial distance calculations are accomplished using Vincenty's formula (Bessel et al., 2010). Loading and pre-processing of the taxi trip records is accomplished using the Dask parallel computation and task scheduling Python library Arabie and Carroll (2016).

Prior to pre-processing, each month of yellow and green taxi trip data amounts to approximately 2.5Gb and 300Mb, respectively, totalling ≈115Gb over the course of the experimental date range. Pre-processing the dataset using a distance criterion $d = 300$ feet reduces this to two files of size ≈20Gb each, one of trips beginning near hotels of interest (*destinations*), and another ending nearby (*origins*).

Pre-processing a month of yellow taxi data using the aforementioned Python process and distance criterion $d = 300$ requires *at most* 40 minutes and 50Gb of random-access memory. Using the super-computing cluster with a 40-node per-user allotment, processing all 42 months of taxi trip date requires only ≈80 minutes. If equipped with more than 42 nodes, the processing time would be cut in half.

3.2 *Distance criterion optimization*

A crucial assumption in the analyses of the taxi dataset is that taxi trips that originate or culminate near a hotel is likely to indicate that a guest is arriving or leaving the hotel. With this in mind, we can use the locations to which the guest travels, or where the guest has traveled from, to draw conclusions about typical patrons per hotel.

The distance criterion d should be selected such that it maximizes economic predictability in some sense. We compare a dataset of occupancy rates with proportions of nearby taxi trips, per hotel. We are interested in finding a setting of d such that some divergence measure $\mathcal{D}$ between empirically observed distributions $\hat{p}_{\text{occupancy}}$ and $\hat{p}_{\text{taxi}}$ is minimized. We consider relative entropy, total variance distance, and total relative difference.

After pre-processing using a suitably large distance criterion (e.g., $d = 300$ft.), a range of candidate distance criterions are considered (e.g., $d \in [25\text{ft.}, 50\text{ft.}, ..., 275\text{ft.}, 300\text{ft.}]$). The

empirical distribution $\hat{p}_{\text{taxi}}$ is computed from the data for each choice of d, and we calculate the selected divergence measure with respect to the fixed $\hat{p}_{\text{occupancy}}$. The criterion d giving the minimal divergence value is said to be the "best", and the dataset obtained from this choice of d can be used for downstream analyses.

3.3 *Removing outlier hotels*

Certain hotels under consideration may have an abnormal large or small number of nearby taxicab pick-ups or drop-offs depending on their location in the city. For example, Hotel Pennsylvania, located adjacent to Pennsylvania station, has a disproportionately large number of both pick-up and drop-off taxi trips due in part to the traffic that the train station incurs. This observation makes it difficult to justify using such hotels in any joint hotel occupancy-taxicab ride distribution analysis.

Using the same divergence measures as listed above, we propose two methods for the removal of outlier hotels:

1. Iteratively remove hotels from the occupancy dataset during the distance optimization process until the divergence measure between distributions, $\mathcal{D}(p_{\text{occupancy}}||p_{\text{taxi}})$, is below some threshold. Picking the threshold is a hyper-parameter choice, and may be difficult to select *a priori*.
2. After running the distance optimization procedure, restore any hotels that may have been removed from the dataset, and apply a outlier-detection algorithm based on the per-hotel divergence measure values.

4. Results

5. Conclusion

Acknowledgements

The authors would like to thank the College of Computer and Information Sciences for the gracious allowance of supercomputing resources.

References

Arabie, P. and Carroll, J. D. (2016). *Dask: Library for dynamic task scheduling.*

Bessel, F. W., Karney, C. F. F., and Deakin, R. E. (2010). The calculation of longitude and latitude from geodesic measurements. *Astronomische Nachrichten* **331,** 852–861. arXiv: 0908.1824.

Chu, D., Sheets, D. A., Zhao, Y., Wu, Y., Yang, J., Zheng, M., and Chen, G. Z. (2014). Visualizing hidden themes of taxi movement with semantic transformation. *2014 IEEE Pacific Visualization Symposium* pages 137–144.

Doraiswamy, H., Ferreira, N., Damoulas, T., Freire, J., and Silva, C. T. (2014). Using topological analysis to support event-guided exploration in urban data. *IEEE Transactions on Visualization and Computer Graphics* **20,** 2634–2643.

Doraiswamy, H., Vo, H. T., Silva, C. T., and Freire, J. (2016). A gpu-based index to support interactive spatio-temporal queries over historical data. *2016 IEEE 32nd International Conference on Data Engineering (ICDE)* pages 1086–1097.

Ferreira, N., Poco, J., Vo, H. T., Freire, J., and Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics* **19,** 2149–2158.

Huang, X., Zhao, Y., Ma, C., Yang, J., Ye, X., and Zhang, C. (2016). Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi

trajectory data. *IEEE Transactions on Visualization and Computer Graphics* **22,** 160–169.

Stoyanovich, J., Gilbride, M., and Moffitt, V. Z. (2017). Zooming in on nyc taxi data with portal. *CoRR* **abs/1709.06176,**.

Wu, F., Wang, H., and Li, Z. (2016). Interpreting traffic dynamics using ubiquitous urban data. In *SIGSPATIAL/GIS*.