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

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. Predictions of hotel occupancy rates are conditioned on relevant taxi records, showing an improvement over a baseline predictive model. We show exploratory visualizations and analysis of taxicab trip records. We suggest and preliminarily investigate applications of the New York City taxicab trip records data 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 Taxi & Limousine Commission 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 more.

Combining the NYC taxicab trip records data with additional relevant datasets may allow for more interesting economic conclusions. We study a unification of the taxi data with the occupancy rates of 178 NYC hotels from 2013 - 2016. We are interested in those taxicab 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 create an optimization procedure motivated from a simple distribution matching procedure. Hotels with abnormal amounts of taxicab traffic are eliminated from the data during 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

TODO.

3. Methods

4. Discussion

Acknowledgements

Supplementary Materials

Web Appendix A, referenced in Section **??**, is available with this paper at the Biometrics website on Wiley Online Library.

References

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Appendix

Title of appendix

Put your short appendix here. Remember, longer appendices are possible when presented as Supplementary Web Material. Please review and follow the journal policy for this material, available under Instructions for Authors at `http://www.biometrics.tibs.org`.