Is there a relationship between complicated crossings and frequently visited locations? – A case study with boro taxis and OSM in NYC

Andreas Keler*, Jukka M. Krisp*

* University of Augsburg, Institute of Geography, Applied Geoinformatics, Alter Postweg 118, 86159 Augsburg, Germany

Extended Abstract

Introduction

Extracted positions of taxi trajectories, where customers enter or leave the vehicle are useful information for a big variety of applications (Castro et al. 2013). They are useful for estimating cruising miles, which is the taxi mileage without a fare, and helpful for its reduction as tested with the pCruise algorithm (Zhang & He 2012). Additionally, the inspection of taxi pick-up points might be helpful for providing taxi service quality assessments (Zhang et al. 2014) and analyzing the taxi drivers behavior, for example if the driver is hunting or waiting for new customers (Li et al. 2011). For the inspection of the spatial distributions of taxi pick-up and drop-off points, Pan et al. (2013) are using besides "heatmaps", which are 2-D histograms in the Euclidean space, as well the density-based clustering algorithm DBSCAN (Ester et al. 1996), which was successfully tested with geodatabases (Sander et al. 1998). In contrast to this, Krisp et al. (2012) propose partitioning-based clustering (k-means) for getting in and out of a taxi. When taxi pick-up and drop-off points in Euclidean space form density- or distance-connected clusters of certain spatial extent, we can associate those areas with additional geographic information on land-use (Pan et al. 2013) or on region function (Li et al. 2011). Density-based clustering is applicable for whole vehicle trajectories as shown by Rinzivillo et al. (2011) with the use of the OPTICS algorithm, which was introduced by Ankerst et al. (1999). The big advantage of OPTICS is the possibility to create reachability plots that allow estimations for beneficial search distances (Epsilon) for a selected minimum number of points (MinPts).

LBS 2016
13th Conference on Location-Based Services
Vienna, 14–16 November 2016

Published in "Proceedings of the 13th International Conference on Location-Based Services", edited by Georg Gartner and Haosheng Huang, LBS 2016, 14-16 November 2016, Vienna, Austria.

In our approach we use OPTICS for extracted taxi drop-off points for the derivation of density-based clusters in one selected investigation area on one selected day of observation. These resulting clusters are then related with complicated crossings.

Description of the test data sets

In this work we inspect the vehicle traffic dynamics and the transportation infrastructure of New York City (NYC). Therefore, we make use of time-based extracts from green taxi trip data and of road network data within the administrative boundaries of NYC coming from the OpenStreetMap (OSM) project.

Trip data of green taxi cabs in New York

New York City, especially Manhattan, has a high coverage of frequently operating taxis. Besides the typical yellow taxis there are also green taxis, which are also called “boro taxis”. Those green taxis were introduced in 2013 for gaining more distribution in the boroughs outside Manhattan. This was initiated after inspecting the distribution of pic-up points of yellow taxi GPS trajectories: 95 % of pic-ups occurred within Manhattan and the rest in the outer boroughs[1].

Our data set is coming from the NYC Taxi & Limousine Commission (TLC)[2] and includes taxi trip records from all trips.

For our approach of detecting frequently visited locations within a short distance radius, we are only using the taxi drop-off points, where customers are leaving the green taxis.

As we suppose, we have more fluctuations from outer boroughs into Manhattan on weekends as on working days. Therefore, we select one Saturday in 2015: 13th of June. In total there are 24.820 green taxi trips with the same number of drop-off points. From the 20 available attributes, we only use geodetic coordinates of drop-off points. There is no given attribute indicating specific taxi identifications.

---

[1] Background on the Boro Taxi program. NYC Taxi & Limousine Commission. URL: http://www.nyc.gov/html/tlc/html/passenger/shl_passenger_background.shtml; Retrieved December 18, 2013.

[2] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

NYC road network information from the OpenStreetMap (OSM) project

OpenStreetMap (OSM) is an open Crowdsourcing project that was founded in 2004 by Steve Coast with the aim to create an open geodatabase. Since then, worldwide volunteered mappers produce Volunteered Geographic Information (VGI). For Stanica et al. (2013) the OSM has one of the most accurate freely available road networks. Its practical use for vehicle routing was tested by Graser et al. (2014) for the case of Vienna. The conversion into 1-D network space sometimes comes along with problems in connectivity and has to be modified for practical routing applications.

Defining complicated crossings and frequently visited locations in NYC

Our Method consists of creating polygons based on two different data sources. Besides complicated crossings from OSM road networks, we infer taxi drop-off point cluster polygons. After matching these two polygon types, we define frequently visited locations. This workflow is pictured in the diagram in *Figure 1*.

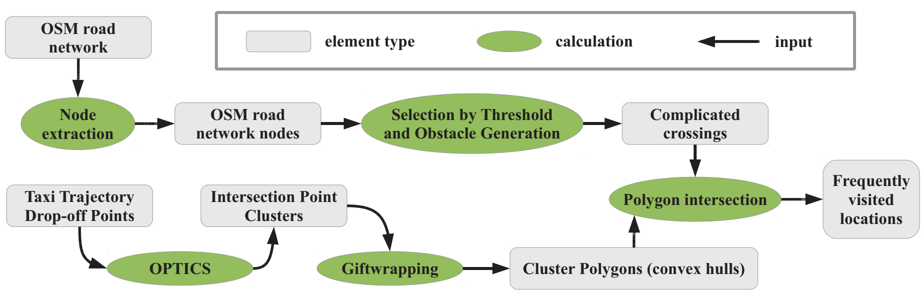

Figure 1. Workflow for the definition of "frequently visited locations".

We have rethought the idea of introducing a representative search distance for applying the OPTICS algorithm, which was introduced by Ankerst et al. (1999).

As provided in Krisp & Keler (2015) we can use an average distance of roundabouts (in this case 60 m) for selecting a threshold for the creation of obstacle polygons. These obstacle polygons are so called complicated crossings and are based on two-dimensional Kernel Density Estimation of street nodes (see *Figure 1*). In case of typical American cities we have to think about something else: Since Summer 2015 there is only one roundabout in New York City. This pictures the unpopularity of this transportation infrastructure element, which is very popular in Western Europe: France has with over 30.000 roundabouts the highest number in the world. After inspecting the transportation infrastructure in Manhattan, we keep the search radius on 60 meters.

In parallel we have to define taxi drop-off point clusters by using the OPTICS algorithm. In case of this algorithm, we have to define a search distance (Epsilon) and a minimum number of points (MinPts) for estimating the density. For the search distance Epsilon we take the maximum Street width in Time Square of 102 feet, which equals around 31 meters (31.0896 meters). Our minimum number of taxi drop-off points will be 2 (MinPts = 2). With those two parameters we define our density connection between the drop-off points. The resulting 1.867 clusters are then transformed by the gift wrapping algorithm (Jarvis 1973) into polygons.

In the last step of defining frequently visited locations we match the 613 complicated crossings with 1.867 taxi drop-off point polygons in NYC. *Figure 2* shows a cutout of the in total 1.466 defined frequently visited locations in NYC on Saturday, June the 13th 2015, which will be the base for our further analyses.

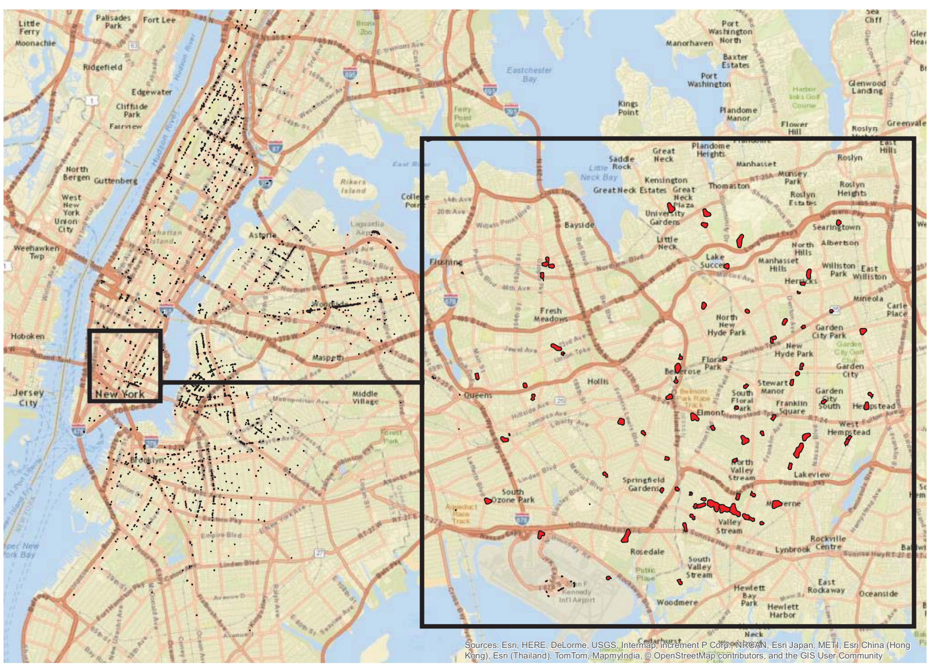

Figure 2. Defined frequently visited locations in NYC on Saturday, June the 13th 2015.

References

Ankerst M, Breunig MM, Kriegel H-P, Sander J (1999) OPTICS: Ordering Points To Identify the Clustering Structure. In: ACM SIGMOD international conference on Management of data. ACM Press, pp. 49-60

Castro PS, Zhang D, Chen C, Li S, Pan G (2013) From taxi GPS traces to social and community dynamics: A survey. ACM Computing Surveys (CSUR), 46(2): Article No. 17

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for sparse representations. In: Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226-231

Graser A, Straub M, Dragaschnig M (2014) Is OSM Good Enough for Vehicle Routing? A Study Comparing Street Networks in Vienna. In: Gartner G, Huang H (eds) Progress in Location-Based Services, Lecture Notes in Geoinformation and Cartography. Berlin: Springer, pp. 3-18

Jarvis RA (1973) On the identification of the convex hull of a finite set of points in the plane. Information Processing Letters 2: 18-21

Krisp JM, Keler A (2015) Car Navigation - Computing Routes That Avoid Complicated Crossings. International Journal of Geographical Information Science (IJGIS) 29(11): 1988-2000

Krisp JM, Polous K, Peters S, Fan H, Meng L (2012) Getting in and out of a taxi: spatio-temporal hotspot analysis for floating taxi data in Shanghai. In: 6th International Symposium "Networks for Mobility", Conceptual paper, September 27/28, 2012.

Li B, Zhang D, Sun L, Chen C, Li S, Qi G, Yang Q (2011) Hunting or Waiting? Discovering Passenger-Finding Strategies from a Large-scale Real-world Taxi Dataset. In: 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), MUCS 2011: 8th IEEE International Workshop on Managing Ubiquitous Communications and Services, Managing Contexts, pp. 63-68

Pan G, Qi G, Wu Z, Zhang D, Li S (2013) Land-Use Classification Using Taxi GPS Traces. IEEE Transactions on Intelligent Transportation Systems 14(1): 113-123

Rinzivillo S, Pedreschi D, Nanni M, Giannotti F, Andrienko N, Andrienko G (2008) Visually driven analysis of movement data by progressive clustering. Information Visualization 7(3-4): 225-239

Sander J, Ester M, Kriegel H-P, Xu X (1998) Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery 2(2): 169-194

Zhang D, He T (2012) pCruise: Reducing Cruising Miles for Taxicab Networks. In: Proceedings of the 2012 IEEE 33rd Real-Time Systems Symposium (RTSS 2012), pp. 85-94

Zhang D-Z, Peng Z-R, Sun DJ (2014) Comprehensive Taxi Assessment Index Using Floating Car Data. Journal of Harbin Institute of Technology 21(1): 7-16