Testing a Route Probability Model with Floating Car Data

David Jonietz[a]
[a] Institute for Geography, University of Augsburg
Alter Postweg 118
86157 Augsburg, Germany
david.jonietz@geo.uni-augsburg.de

Carolin von Groote-Bidlingmaier[a]
cvgb@geo.uni-augsburg.de

Sabine Timpf[a]
sabine.timpf@geo.uni-augsburg.de

Abstract

Negative effects of motorized transportation are often locally concentrated in the vicinity of certain destinations, such as airports, malls or train stations. For transportation planning, modeling the access routes to these locations is particularly challenging, mainly because the origins of the car drivers are unknown. In this study, probable access routes of taxi drivers to a train station are modeled, and compared to actual routes obtained from floating car data (FCD). It is shown that while in several cases, relative route frequencies can be predicted in the vicinity of the destination, others were either missed or inaccurately included in our results. Possible reasons and improvements to the model are discussed.

Keywords: route probability, network analysis, floating car data

1 Introduction

A particular problem for urban transportation planners is to cope with local concentrations of high traffic volumes, which may lead to negative effects such as traffic jams, reduced air quality or shortages of parking space [7]. In many cases, these congestions develop in the vicinity of particular destinations, which attract a high share of motorized traffic, such as main shopping districts, airports or train stations. In order to improve such situations by means of precedent planning or managing transport with intelligent transportation systems (ITS), the particular challenge lies in predicting probable access routes of drivers to the particular destination whose exact point of origin, however, is unknown or at least uncertain. In some cases, the origin of the trip can be approximately described as somewhere inside a given polygon, as for instance within a traffic analysis zone (TAZ), a fundamental spatial unit used for travel demand modelling, or within an administrative area [5, 9].

Concerning both examples, the focus is on the identification of the road segments which are most probably visited by drivers who start their trip from an uncertain location somewhere inside a given polygon. Previously, a method has been proposed which involved using a large set of candidate points as potential origins, each of which is taken as the input for a shortest path calculation to a fixed destination point outside of the polygon [9].

Based on these insights, this study is motivated as follows. Firstly, due to a lack of empirical data, the results of our previous study could not be validated. The availability of large sets of floating car data (FCD), especially obtained from taxi drivers, however, allows for a comparison of calculated frequencies with actual empirical data. Secondly, while [9] applied a simple shortest path algorithm, the routing algorithm in this work will factor in insights about the actual route choice heuristics used by taxi drivers, as described in the literature [8, 10]. Thus, the aim of this study is to apply a similar method to a scenario for which such FCD exists, apply a more elaborate routing algorithm, and present a methodology for model validation.

This paper is organized as follows: First, the method used for route calculation and the floating car dataset will be described. In the following, our method will be discussed before the results are presented. Finally, we provide a discussion and propose future research activities.

2 Calculating Route Probability from an Uncertain Origin to a Destination

In our previous study, the aim was to compute the probability of visitors of a football game choosing specific routes from their home county to the stadium [9]. Thus, while their destination is exactly known, their point of origin can only approximately be described as within an administrative area. Representing the road network as a graph $G = (V, E)$, the probability $p(e)$ for each edge $e \in E$ being visited by a visitor is calculated. Following a frequency-based approach, $p(e)$ of this event can be inferred from its relative frequency $f(e)$ in a number of trials, so that $f(E) = p(E)$ [3].

Using a range of approaches, including purely geometrical and geographically-weighted methods, large sets of candidate points $V' \in V$ located within the polygonal boundary, which represents the potential area for the starting location, are calculated. Then, shortest paths are computed for each pair of candidate point and destination. Afterwards, the route frequency $f(e_{1\ldots n})$ for each edge e is obtained based on the number of overlapping routes. In order to infer $p(e_{1\ldots n})$ from $f(e_{1\ldots n})$, the results are normalized by division of the frequency value for each segment i with the maximum frequency value f_{max} found among all network segments. These values, which range from 0 - 1 for each are called *R-values* in [9].

Comparing the results, it could be concluded that the different methods of conceiving the set of candidate points produce only slightly different results, so a geometric solution, which can be more feasible in terms of computational effort or data requirements, may be sufficient [9].

3 Floating Car Data

In this study, building on the results of [9], a model of route probability calculation will be compared to actual route choice behavior as obtained from FCD. Today, there are several datasets available which include tracking data of fleets of commercial but also private vehicles. These data are generally obtained from GPS-based positioning and include x, y, z coordinate tuples which are automatically mapped at a predefined certain time interval. In the past, FCD have been used in a variety of studies, including as a basis for dynamic navigation systems [8], in order to estimate traffic conditions [1] and develop route choice models for heavy good vehicles [4], commuters [2], or taxi drivers [10]. For the latter, it has been found that they tend to optimize travel speed, minimize the number of left turns and prefer roads of a higher hierarchical level, such as expressways [8, 10].

There are, however, several problems and potential pitfalls when using FCD, including privacy issues but also errors or inaccuracies which might require extensive post-processing of the raw data [2].

Concerning the dataset used in this study, it was collected by [6] from approximately 500 taxis over 30 days in the San Francisco Bay Area and includes, apart from the temporal and location data, further information such as the taxi ID and the status as free or occupied. Due to positional errors, occasional temporal and spatial gaps between tracked points and attributive inaccuracies, especially concerning the occupation status, several processing steps were necessary before using the data for analysis, which will be described in the following chapter.

4 Method

In this section, our method is described, starting with the routing and probability calculation. In the following, the focus is on the post-processing of the FCD and the comparison of the results.

4.1 Calculating Route Probability

In accordance with the FC dataset of [6], San Francisco (California) was chosen as study area. As a distinct destination point v_d for our study, we choose the San Francisco Caltrain Station as a representative location which might attract a high share of motorized traffic, so that resulting problems are to be expected in its vicinity. In addition, a preliminary check of the FC dataset showed that there are a sufficient number of GPS trajectories which terminate there. As a polygonal area P to represent the uncertain starting point, the administrative zip codes 94105, 94110, 94117, 94123 were chosen so that $v_d \notin P$.

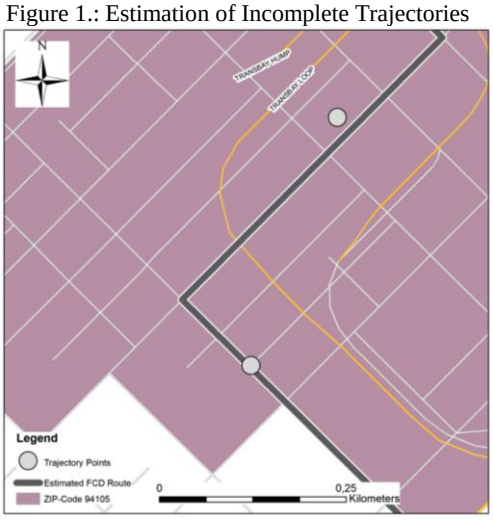
In a first step, following the argumentation by [9], a simple point generation method was applied to create a set of candidate points V' within the zip code polygon P. Thus, 1000 points were created and randomly dispersed among the road network so that each $v' \in V$ and $v' \in P$. Then, optimal paths were calculated for each pair of v' and v_d. According to the typical route choice heuristics of taxi drivers described previously, the optimization algorithm preferred faster rather than shorter routes and avoided left turns.

In the following, the route frequency $f(e_{1\ldots n})$ for each edge e could be calculated by counting the number of overlapping routes. In order to infer $p(e_{1\ldots n})$ from $f(e_{1\ldots n})$, the results were normalized by division of the frequency values with the maximum frequency value f_{max} found among all network segments in order to receive *R-values* between 0 – 1.

4.2 Validation with FCD

Before using the FCD for further analysis, several steps of processing were necessary. As a first step, only the trajectories were selected, which changed their status from unoccupied to occupied within the selected zip code area, and turned again to unoccupied in the direct vicinity, in our case a buffer distance of 50 m, of the San Francisco Caltrain Station. Thus, we received 729 trajectories in total, which originated within P and terminated at v_d. In the next step, the tracking points were snapped to the nearest road network segment. Unfortunately, the temporal intervals between recorded locations were not constant, but in some cases, there were longer intervals with no recorded coordinate pairs. In these cases, the fastest path was calculated in between the remaining points in order to achieve probable routes, as shown in figure 1. As illustrated, this method may lead to inaccuracies, since the actual route between these two recorded points is not known.

Figure 1.: Estimation of Incomplete Trajectories



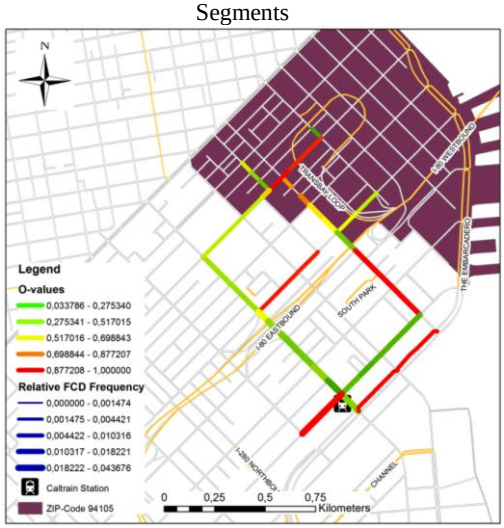
Finally, visual analysis of the set of trajectories demonstrated that a certain number clearly included more than just one trip, but did not change the occupation status attribute accordingly. In order to cope with these inaccuracies, for each trajectory, the according shortest path was calculated and the lengths compared to each other. Trajectories which were longer than the mean of the relative difference were excluded from further analysis. In the following steps, the absolute and relative route frequency was calculated for each road network segment as described previously, and the difference calculated among the

simulated values and the trajectory-based values. For this reason, both *R-values* were normalized by division by the larger one of the two values, and the difference calculated between them. Thus, we receive what we call *O-values*, which can range from 0 (perfect prediction) to 1 (false prediction).

5 Results

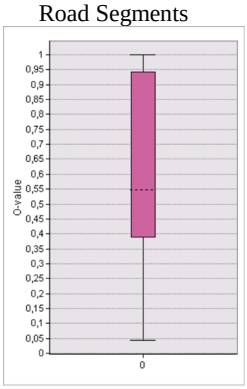
Figure 2 illustrates the resulting *O-values* for a selection of road segments, namely the 50 most frequented as received from the FCD. As one can see, these are mainly in the direct vicinity of v_d, which is not surprising since the incoming traffic from different directions is gradually being channelled there. On the map, it is also visible that there are huge differences regarding the predictive quality of the model, with relative frequencies on some road segments being predicted to a high degree while others were either missed or inaccurately included in our routing.

Figure 2.: Results for the 50 Most Frequented Road Segments



The overall quality of the model is more clearly visible in figure 3, which shows a boxplot of the *O-values* for the respective road segments. As one can see, the mean of *O-values* lies at 0.55, with a large variability of results. The lower and upper quartiles lie at 0.39 and 0.94, respectively.

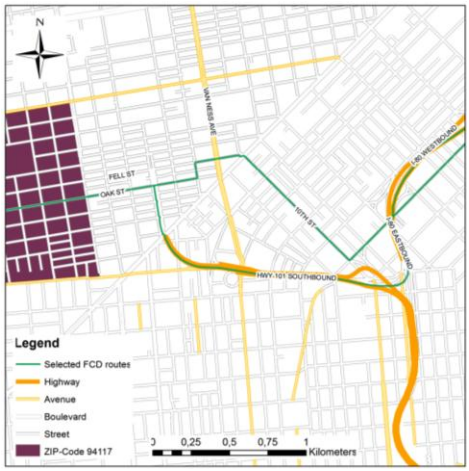
Figure 3.: Boxplot of O-Values for the 50 Most Frequented Road Segments



6 Discussion of the Results

In this chapter, the model results will be discussed. Although several highly frequented road segments, especially near the destination point, were predicted accurately, there are also some cases in which the predicted routes differed to a high degree from the ones obtained from the FCD. Some cases are rather obvious, such as the road segment directly south-west of the destination point, as can be seen in figure 2. The high frequency value for the taxis can be explained by the presence of a taxi parking area on this side of the station, while in our model, the main entrance, which is located on the north-eastern side of the building, was used as v_d. Representing the destination with more than just one v_d may therefore improve the model quality. Other examples of low predictive accuracy, however, demonstrate the boundaries of our behavioural assumptions, such as perfect spatial knowledge, or the homo economicus concept. As can be seen in figure 4, for instance, of around 75 taxi drivers leaving the polygon on the same road, only one third chose the highway, which, based on the literature on taxi drivers' route choice heuristics, would be preferred in our model, while two thirds actually avoided the highway. Such behavioural differences might also be time-dependent and require a more elaborate routing algorithm.

Figure 4.: Differences in Route Choice Strategies among Taxi Drivers



7 Conclusion

This study aimed to validate a route probability model with FCD. Using the example of taxis approaching an inner-city train station from unknown starting locations, probable routes were calculated and the results compared to actual routes obtained from a FC dataset. As the results have demonstrated, the aim of identifying the probable access routes of drivers to the particular destination has partly been achieved. However, while several access routes have been predicted to a satisfying degree, others were missed or wrongly assigned traffic volume. Thus, it seems as if the general assumptions about taxi drivers' route choice heuristics, on which this study was based, need to be reconsidered, such as highway preference. It is also thinkable that the inclusion of dynamic traffic conditions would be a possible way to improve the model. A

difference to the original study by [9] arises from the network structure. While in the first study, the environmental setting was rather rural, with a few dispersed town centres, in this work we focused on an urban area with a regular street pattern. As a result, differences in model quality can be expected. For future work, therefore, it is planned to examine the effect of network structure, especially network connectivity, on the predictive quality of the model.

References

[1] J. Aslam, S. Lim, X. Pan, D. Rus. City Scale Traffic Estimation from a Roving Sensor Network. In *SenSys'12 Proceedings*, 2012.

[2] N. Dhakar. Route Choice Modeling Using GPS Data. Dissertation, University of Florida, 2012.

[3] A. Hajek. Interpretations of Probability. In: Zalta EN (ed) The stanford encyclopedia of philosophy, 2012. http://plato.stanford.edu/entries/probability-interpret/#FreInt. Accessed 30 November 2014

[4] S. Hess, M. Quddus, N. Rieser-Schüssler, A. Daly. Developing advanced route choice models for heavy goods vehicles using GPS data. In *TRB 93rd Annual Meeting Compendium of Papers*, 2014.

[5] M. G. MacNally. The four step model. In D. A. Hensher, K. J. Button, editors, *Handbook of transport modeling*. Elsevier, Oxford, 2007.

[6] M. Piorkowski, N. Sarafijanovic-Djukic, M. Grossglauser. {CRAWDAD} data set epfl/mobility (v. 2009-02-24). 2009 Downloaded from http://crawdad.org/epfl/mobility/, Nov., 2014

[7] H. J. Miller, S.-L. –Shaw. Geographic information systems for transportation: principles and applications. Oxford University Press, Oxford, 2001.

[8] L. Tang, Q. Li, X. Chang, S. Shaw, Z. Zhao. Modeling of taxi drivers' experience for routing applications. *Science China Technological Sciences* 53(3): 44-51, 2010.

[9] C. von Groote-Bidlingmaier, D. Jonietz, S. Timpf. Calculating Route Probability from Uncertain Origins to a Destination. In G. Gartner, H. Huang, editors, *Progress in Location-Based Services 2014*, 19-32, 2014.

[10] E. Yao, L. Pan, Y. Yang, Y. Zhang. Taxi Driver's Route Choice Behavior Analysis Based on Floating Car Data. *Applied Mechanics and Materials* 361: 2036-2039, 2013.