Using FCD-data for measuring route-choice probabilities

Jürgen Mikat *, Kai-Uwe Thiessenhusen, Peter Wagner
German Aerospace Center (DLR), Institute of Transport Research †
Rutherfordstraße 2, 12489 Berlin, Germany

The data sampled within a large FCD-project with taxi drivers can be used to check the route-choice habits of a community of fairly well-informed people. Not surprisingly, the drivers do not always choose the fastest route, although there are changes in the route-choice under varying traffic loads. To understand, however, the motivations into more detail leading to a particular route choice appears to be far from trivial. Some experiences with this unique data set, that is publicly available, will be reported in this contribution.

1. Introduction

The complexity of traffic in many ways is determined by the individual preferences of its users. Today, mobility for people is a factor strongly incorporated in their personal living plans. For those using own vehicles in order to reach different locations probably a gain of time plays the most important role. Decisions have to be made on the best route, i.e. the length and the travel time needed for a given origin-destination relation. But, the knowledge on the current state of the traffic network is limited, even if street maps are provided and driving experience is assumed. During the day, traffic situations may vary due to the change of load and routes which perform good in the morning can be of disadvantage in the afternoon. Further on, variations may occur between different days of the week and in dependence of the time of year. Besides of these factors drivers have to take into account in order to obtain a good performance of their route, personal preferences as shortest route, fastest route, price, mean speed and comfort may be influencing factors for the route choice.

Route choice today plays also an important role for traffic management systems. Traffic management systems often assume, that a mere provision of information on problems in the street network are sufficient for the drivers to adapt their route choice. However, with insufficient specification of the information several consequences may follow. A blockade of the original route will force all drivers to choose other routes, blocking them, too. Sometimes, it might be better to stay on the original route since most of the drivers are changing routes. In turn, this implies that the way of provision, the amount and the degree of detail of information is crucial to successful traffic management. Also, it is crucial to understand, how and when drivers are following the information provided.

In this paper, a method is demonstrated to get better insight into the route choice behavior of experienced drivers. The method is based on historical data and assumes, that different routes may be accounted for a specific origin-destination pair. These routes may differ in their parameter, and are subject to daily or hourly changes forcing drivers to a new decision on their choice of route. A special group of drivers was chosen which shows some homogeneity in its behavior. Taxi drivers offer the opportunity to study route choice behavior of fairly informed travelers with high experience. They constantly adapt their knowledge based on their daily business, and often choose the same routes for a given relation. A high frequency for a the relation, long-term observations and the choice of a specific group of drivers allows a statistical analysis from historical data. The aim is to find characteristics which demonstrate the route choice behavior provided a high information grade, and how other parameters may influence route choice.

2. Data basis and analysis

For the investigation of route choices taxis in Berlin are regarded. A fleet of about 300 taxis provides data on their current position, given in GPS data, their passenger status (e.g. empty, occupied), together with time and a taxi identification number. The data is collected and integrated in a constantly growing historical data base[1,2]. Up to date, more than 4 million data sets were accumulated. The historical data base allows statistical analysis on the taxis'

* Tel.: +49 (30) 67055-278
Fax: +49 (30) 67055-202
email: [juergen.mikat, kai-uwe.thiessenhusen, peter.wagner]@dlr.de
† http://www.dlr.de/Berlin, http://ivf.dlr.de

preferred routes. Figure 1 shows a plot of taxi data positions merged into a coordinate system with the center of Berlin as origin. The outline of the street map of Berlin can easily be spotted. This picture results from the single measurement points of taxis. The position of the points depends on the time interval of the data logging device, i.e. the combination of the GPS receiver with the data logging rate prescribed by the taxi company. This rate is set according to the information demand of the company, and is strongly dependent on the passenger status. The combination of clock speed, vehicle speed and route choice results in a distribution of single measurement points along the streets of Berlin, forming a data map patterning the real city map. The point density then is a measure for the frequency for a certain street link used by taxi drivers.

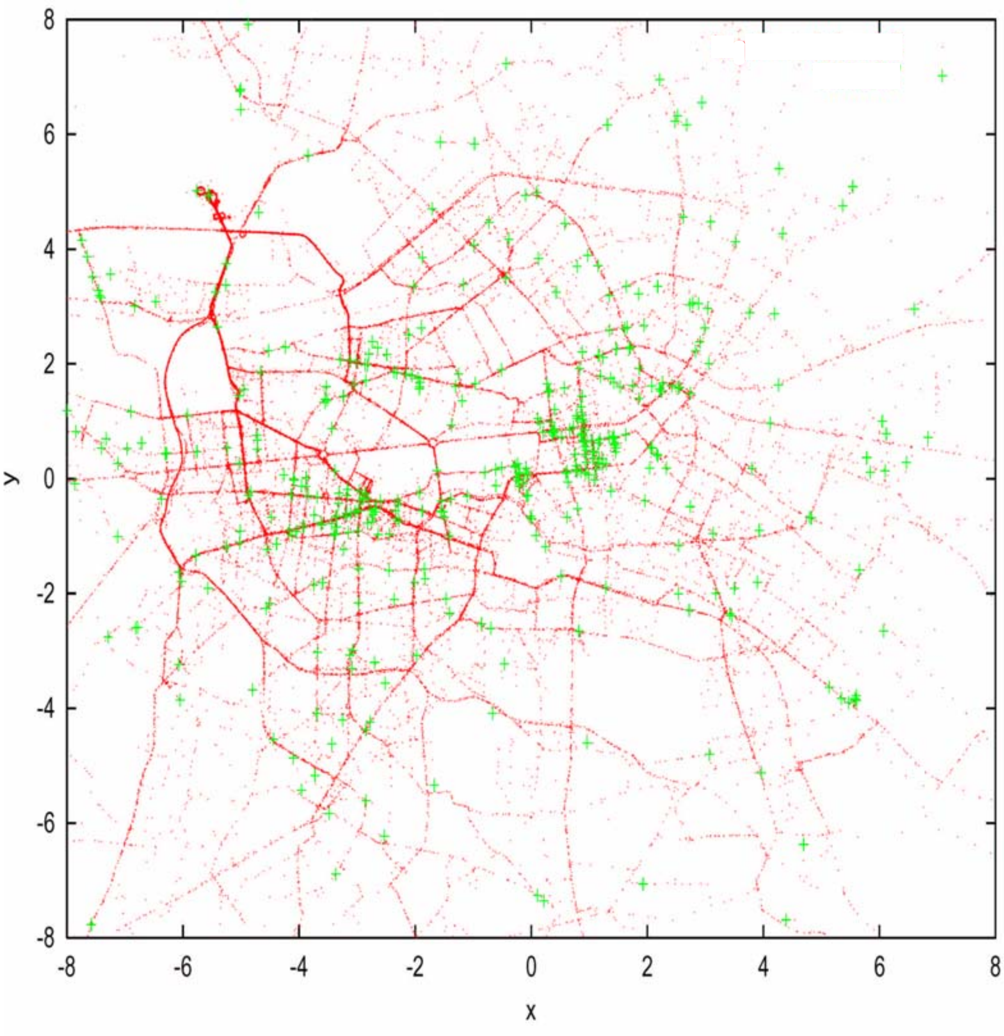

Figure 1: Distribution of taxi positions throughout the city of Berlin as obtained from GPS data collected in an historical data base. Crosses refer to destinations of occupied taxis started from Airport Berlin-Tegel.

For the analysis of route choice behavior an origin and destination may be chosen which gives the most promising relation in terms of statistical significance and reliability of historical data. From the data map areas can be derived with exceptional high point density compared to other streets or areas. These spots can be viewed as points of interest either for the drivers (taxi stands, company headquarter) or the passengers. Most likely, they are preferred areas where taxis can pick up or deliver passengers. Such points may be city centers, airports, train stations, malls etc. One exceptional point in the data map can be found in the north-western part of the city, where the point density is incomparable high. The double loop is the data map representation of the Airport Berlin-Tegel. The Airport offers a unique possibility for the investigations of route choice, since it has a defined departure and arriving area for the taxis, and is a source of customers.

Included in Figure 1 is the distribution of the destination of taxis which are occupied with customers and started out from Tegel airport. Several areas with high point density can be accounted for. An area marked by a high point concentration within relatively few streets can be found east of the Brandenburg Gate, in the zone around Friedrichstraße at position (1,1). About 1/6 of all occupied taxis aim to this destination, therefore providing a most promising relation for route choice analysis.
More parameters may be chosen for route choice analysis, to enhance the quality of the results. First, the boundary conditions along with the origin-destination relation should be identical for all taxis. The extrinsic taxi attributes should therefore be identical, i.e. taxis leaving Airport Tegel should be distinguished between occupied and unoccupied vehicles, as done in Figure 2. Second, parameters influencing the route choice should be identified, i.e. travel time, travel distance, number of stops, average velocity etc. These parameters are discussed in the following section.

3. Route Characteristics for City Center Rides from Airport

Figure 2 shows the distribution of taxis with the origin-destination pair Airport Tegel-Friedrichstraße. Four different main routes can be identified, which are partly overlapping. The routes are denoted by “Hansaplatz”, “Alt Moabit”, “Siemensstraße” and “Otto-Suhr-Allee” according to main spots or streets where the routes are passing by. The route “Otto-Suhr-Allee” is distinguished by the others by including the city freeway for a short piece of the way. The routes, except for “Otto-Suhr-Allee”, differ not much in the total distance. From the taxi data, “Otto-Suhr-Allee” comprises the longest average travel distance with 9.1±0.4 km, followed by “Hansaplatz” (8.1±0.4 km), “Alt Moabit” (7.9±0.5 km) and “Siemensstraße” (7.6±0.6 km). Therefore, within the errors, the latter routes can be considered as of the same length.

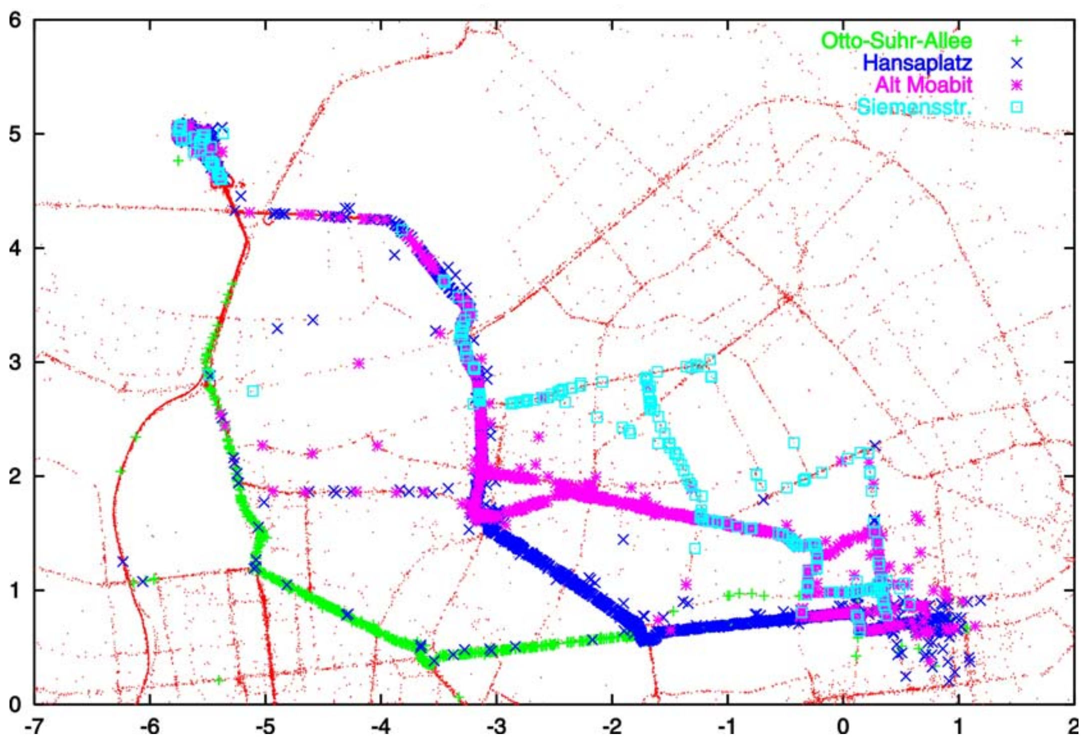

Figure 2: Routes between Airport Tegel and Friedrichstraße taken by occupied taxis.

The taxis taken into account for the analysis were carefully chosen by restrictive filtering of the data. Data only was considered which fulfilled the following conditions: 1. A tour must start at Tegel Airport taxi pick-up area (first data point of tour), and end within a square of 600 m length with the corner Friedrichstraße/Unter den Linden as center (last data point before change of status). 2. A minimum amount of data points is required (at least five). 3. The taxi must be of status “occupied”. 4. Only tours from comparable weekdays were considered, i.e. from Tuesday to

Thursday, holidays excluded. According to this rules, more than 1000 taxi tours could be filtered. The resulting spread of the taxis on the possible routes is shown in Figure 3a, separated into hours of one day.
For most of the taxi drivers the route via "Hansaplatz" is preferable. In the daily mean, 47% of all taxis use this route. In decreasing order, "Alt Moabit" (26%), "Otto-Suhr-Allee" (22%) and other routes (5%), including "Siemensstraße", are following. However, the distribution of taxis on the routes vary during the day. Between 2 pm and 4 pm, "Otto-Suhr-Allee" is preferred, and during the morning and in the evening hours, "Hansaplatz" dominates in the distribution. The strong decrease of the total number of taxis in the early afternoon is due to the characteristics of the origin-destination relation. Most of the taxis carry business people which have appointments in the morning or around noon, resulting in a high peak in the morning hours. In the afternoon, the number of occupied taxis leaving Airport Tegel are increasing again probably caused by business people returning or with businesses in the early morning of the next day.
One might suspect, that the change in route choice behavior is due to a better performance of specific routes during the day. From Figure 3b, it can be seen that in the morning hours the differences in travel times for "Hansaplatz" and "Otto-Suhr-Allee" are low, whereas the travel time of "Alt Moabit" is constantly larger. In the afternoon hours, the travel time needed on the "Hansaplatz" route is drastically increased, compared to the morning hours. It decreases again in late afternoon, where the travel times on all routes are of comparable length (T>5 pm). "Siemensstraße" showed the longest travel time with the largest standard deviation.

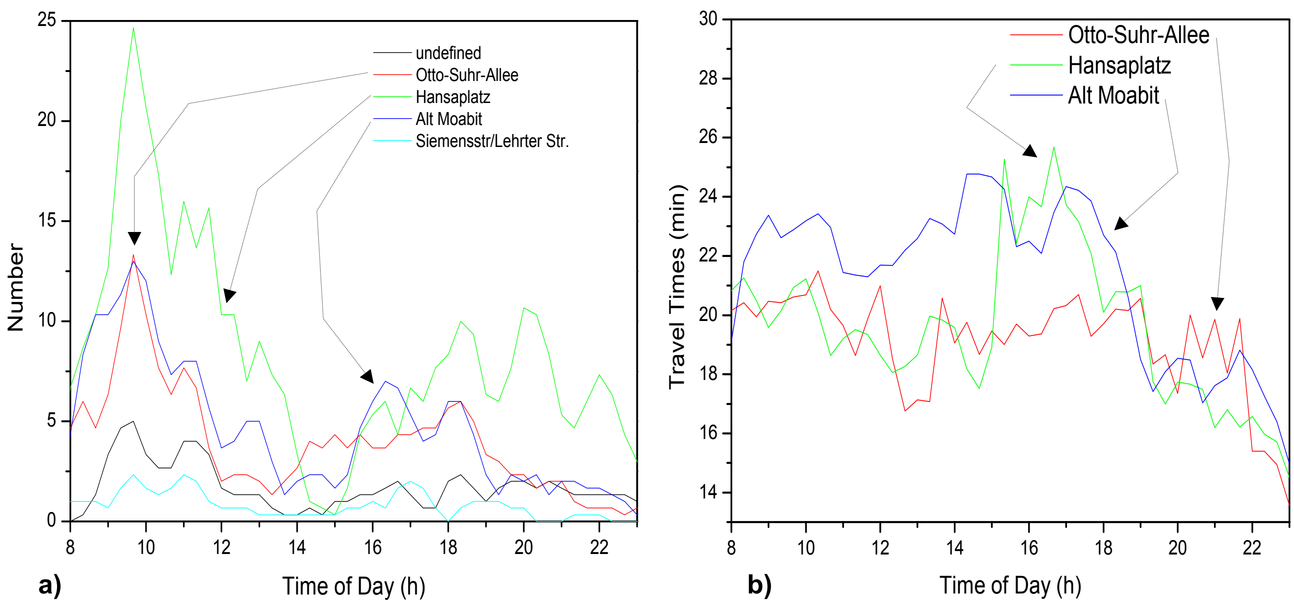

Figure 3: a) Numbers of occupied taxis using different routes between Airport Tegel and Friedrichstraße (20 min averages); b) Travel times needed for main routes (averaged)

The average speed on each route may give another hint for route choice behavior. Besides variations during the day, "Otto-Suhr-Allee" exhibits the largest average speed (29 km/h), followed by "Hansaplatz" (26 km/h) and "Alt Moabit" (23 km/h), with standard variations of about 5 km/h.
Table 1 summarizes the daily averages for all routes. Additionally, the taxi fee gained by the driver is listed.

Table 1: Calculated averages for travel distance, travel time, speed for each considered route between Airport Tegel and Friedrichstraße, and relative numbers of taxis using these routes. Where useful, averages for morning and afternoon hours are also given.

Daily Averages	Routes							
	"Otto-Suhr-Allee"		"Hansaplatz"		"Alt Moabit"		„Siemensstraße"	
Travel Distance [min]	9.1 ±0.4		8.1 ±0.4		7.9 ±0.5		7.6 ±0.6	
Travel Time (am/pm) [min]	19.6 ±1.4	19.5 ±0.9	19.7 ±1.0	20.1 ±3.1	22.3 ±1.1	20.5 ±2.7	26.2 ±4.2	22.9 ±4.9
Speed [km/h]	29 ±5.3		26 ±5.0		23 ±5.3		(25 ±4.7)	
Number of Taxis (total)	22 %		47 %		26 %		5 %	
Number of Taxis (am/pm)	21 %	22 %	48 %	47 %	27 %	25 %	4 %	6 %
Taxi Fee	16,70 €		15,60 €		15,30 €		15,60 €	

4. Route Choice Analysis

Data analysis

From the data shown above, taxi drivers transporting customers decide upon their route almost independently from the travel distance. The shortest route is taken into account only from a minority of the drivers during the day. There is a clear preference for a route with a "medium" travel distance, whereas the route with the longest distance is by far not neglected, but chosen by about one fifth of the drivers. However, from the travel times, no difference between the longest route and the most frequently used route can be detected for the morning hours, and the route with comparable length to the most frequently used is exhibiting a longer travel time. Again, nevertheless, this route is chosen by one quarter of the drivers. This behavior is not altered in the afternoon. Now, the longest route is also the quickest in average ("Otto-Suhr-Allee"). But, the amount of drivers preferring this route is almost unchanged (except for a short period in the early afternoon, where the travel time is strongly increased for "Hansaplatz" and therefore a switch towards the "Otto-Suhr-Allee" is almost trivial). Instead, drivers stick to "Hansaplatz", which is now comparable to "Alt Moabit". On the latter route, the travel time in the afternoon is lowered of about two minutes, but a smaller part of taxis than in the morning choose this route. Some of them may now use "Otto-Suhr-Allee", or even may switch to "Siemensstraße" which still exhibits the largest travel time.

As a hypothesis, one can deduce from this data that taxi drivers look at the travel time in combination with the travel distance. Long travel times seems rarely to be acceptable for even short travel distances, where long travel distances seem to be acceptable when provided with a short travel time. If the obtained values for Travel Time and Number of Vehicles on the corresponding route are grouped into long, medium and short distances and times, according to the actual data referred to for comparison, Table 2 may be obtained for visualization of this relationship.

Apparently, there is a discrepancy in the medium travel time range. Despite of the acceptance of the medium travel time during the morning hours, routes with comparable travel time are almost neglected in the afternoon. The reason for this behavior may be twofold. First, the standard deviation shows a highly untrustworthy average travel time for "Siemensstraße". Therefore, if a risk of large delay may persist on the route, this route may be unfavorable. Second, a route which performs good in the morning may leave the drivers attached to it in the afternoon, too, i.e. taxi drivers may show a moment of inertia for route changes.

Table 2: Preferences of taxi drivers for combinations of travel times and travel distances. Denotation for travel distances: long - "Otto-Suhr-Allee", short - "Hansaplatz", "Alt Moabit" and "Siemensstraße"; travel times: short – 19.5-20.5 min, medium – 22-23 min, long - >26 min (missing values account for a non-existing travel time/travel distance relationship).

Travel Distance	Travel Time		
	long	medium	short
am long	-	-	21 %
short	4 %	27 %	48 %
pm long	-	-	22 %
short	-	6 %	72 %

Correlation analysis

In order to get more insight into this behavior, the data was investigated in detail for the 24 hours of a day. The correlation between the number of vehicles and the travel time needed on a route is calculated for each hour of the day for all routes except for "Siemensstraße" which portion in total number of vehicles is too small. The corresponding plots are shown in Figure 4 for each route in relative number of vehicles. For these plots the early morning hours and late evening hours were neglected, since the travel times were unreasonable low. In an ideal case, an anti-correlation can be expected, i.e. on a specific route the number of vehicles on this route should increase with decreasing travel time. This approach assumes that a route is preferentially chosen if its travel time decreases, and a route with increasing travel time is rather neglected. Note that the approach incorporates a mutual dependency of the plots, since the routes compete with each other.

The plots indicate that the drivers anticipate the changes in travel time on "Hansaplatz" quite well. For the route "Otto-Suhr-Allee" a tendency for, and for "Alt Moabit" a good regular correlation can be found. This unexpected correlation for Alt Moabit, where longer travel times are better accepted than shorter ones can be understood when compared to "Hansaplatz". This behavior is perspicuous, if "Alt Moabit" generally exhibits longer travel times, but is preferred when "Hansaplatz" becomes worse. A similar, but less strong behavior can be found for "Otto-Suhr-Allee". The correlation was calculated using the standard Bravais-Pearson-Korrelation,

$$r(x,y) = \frac{1}{n} \cdot \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \frac{\mathrm{cov}(x,y)}{s_x \cdot s_y}$$

with, in this case, x_i are the travel times for each hour of the day, y_i the respective number of vehicles having the travel time x_i , n the total number of vehicles, $\bar{x}$ and $\bar{y}$ the averages of all x and y, respectively, and s the standard deviations of the respective values.

The only slightly positive coefficient results in a determination coefficient of r^2=0.05 indicating that there is almost no correlation between the number of vehicles on the route and the travel time and their respective variances. For "Hansaplatz" and "Alt Moabit" are r^2=0.34 and r^2=0.39, respectively.

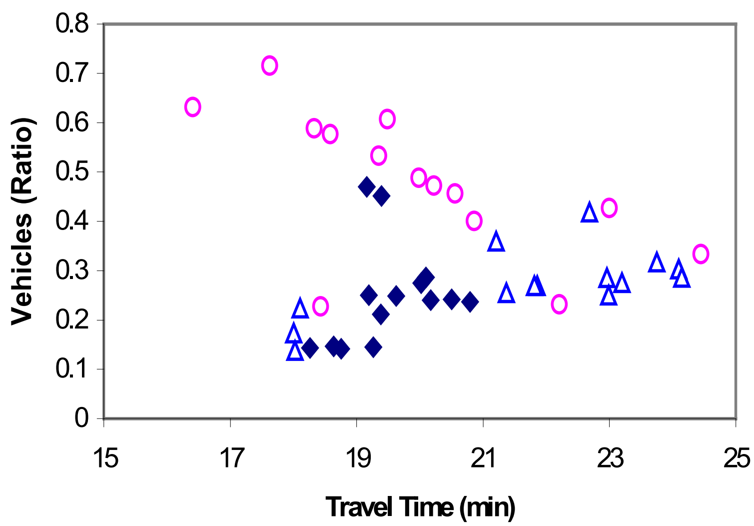

Figure 4: Correlation between the relative amount of vehicles on each route and the travel time on the same route. Symbols and correlation coefficients *r* : ♦: "Otto-Suhr-Allee" (*r* =0.23), O: "Hansaplatz" (*r* = -0.58), Δ : "Alt Moabit" (*r* =0.62)

A measure for the excellence of the drivers route choice behavior can be deduced by considering the portion of taxis using the route with the shortest travel time. In Figure 5 this correlation is shown for each hour of the day. Additionally, the total number of taxis is shown to give some hint on the quality of the calculated correlation. The strong variation of the correlation coefficient during the day indicates a sub-optimal knowledge of the route with the shortest travel time, or the deliberately preference of other routes by the taxi drivers.

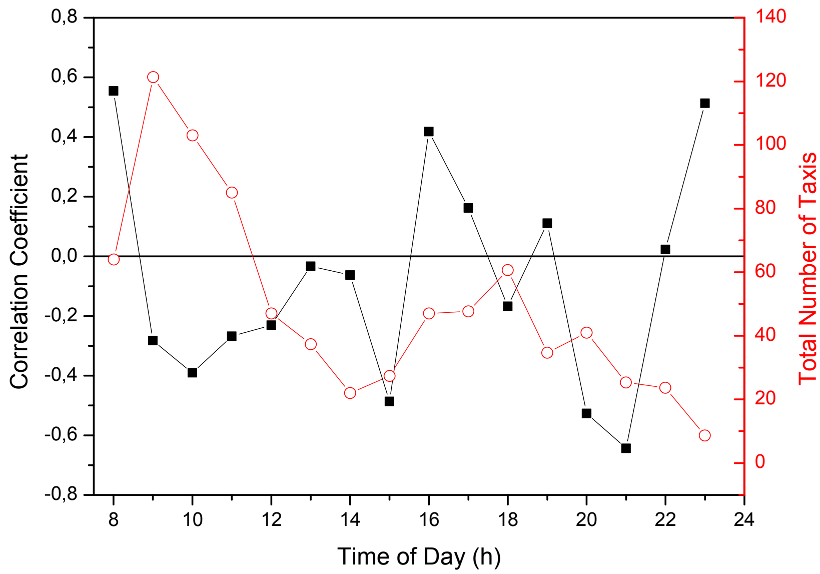

Figure 5: Correlation of taxi route choice with the route of shortest travel time. Left axes: correlation coefficient (■); right axes: total number of taxis per hour (O)

Summarizing, the change of the travel times on the route "Hansaplatz" during the day seems to be noticeably clear to the drivers, whereas the lowering of the travel time on "Siemensstraße" is not perceived, probably due to the large standard deviation.

Subjective decision factors with possible impact

Besides the apparent factors of travel time and travel distance, other factors less obvious may also determine the route choice behavior, e.g. the average speed, the speed sensed by the drivers or the number of stops on the route.

Finally, some of those possible reasons with effect on the route choice should be regarded. First, the **fee** gained from the customers may play a decisive role for the route choice. Since "Otto-Suhr-Allee" is the longest route, and according to the taxi fee regulations in Berlin the charge for a ride is primarily proportional to the distance traveled (a base fee is added for each ride independent of the route), taxi drivers would gain most profit from this route. However, about 75 % of all drivers choose the other routes which is in favor of the customers purse.

On the other hand, the routes exhibits different stabilities with respect to the travel times: "Otto-Suhr-Allee" has the lowest variability during the day, "Hansaplatz" and "Alt Moabit", both, show large differences in the variability between morning hours and afternoon. This **robustness**, eventually combined with the relatively high speed possible on this route, may be the reason for one quarter of taxi drivers to choose "Otto-Suhr-Allee" even in the morning hours. The similar, but vice versa, argument may be accounted for "Siemensstraße", which shows the highest unreliability. In the afternoon, a switch from "Hansaplatz" to "Alt Moabit" and back again takes place, according to the travel times. Nevertheless, the question remains, what reasons lie behind the preference of "Hansaplatz" even when "Alt Moabit" shows not only the same distance but also the same travel time (evening hours).

As additional information we considered the **"smoothness"** of a route as a parameter. With "smoothness", a subjective perception is meant, which is triggered by the number of stops on a route, the waiting time at the stop lights, the speed between the stops and its variation. For example, continuous driving with only few stops may lead to the impression of a route with short travel time. If continuous driving is combined with low average speed, an opposed effect may occur, a route may be sensed as less diversified and boring. A couple of stops combined with high speed on each link between the stops may be perceived as a quick route, even if the total travel time is larger than on a boring route.

As an approach to quantify "smoothness", the number of stops and the average speed per link is considered. The number of stops cannot be derived directly, since it does not necessarily correspond to the number of intersections or traffic signals. For example, a coordination of traffic signals yields a low number of stops but a high number of intersections crossed. In order to obtain a measure for the number of stops, single data points of the tours between Airport Tegel and Friedrichstraße are observed for each route. The results are given in Figure 6. Apparently, clusters of data points are formed at certain distances from the Tegel Airport, sometimes separated by large gaps. Since the data points are gathered with a certain time rate (3 min), clusters corresponds to distances where lower speed or stops are imposed on the taxis, i.e. the distance between two measurements is small. Clustering occurs, when the vehicles approach intersections, ramps, roundabouts, or drive in congested areas. Vice versa, gaps between the clusters indicate route sections, where the taxis are able to drive with higher speed. Before and after noon, some differences may occur due to different vehicle flows. High vehicle densities may decrease the average speed and therefore blur the relation between measurement points and stops. On the other hand, an even distribution of the points may indicate a constantly slow speed, where stops can no longer be identified.

For "Hansaplatz", four broad clusters can be identified, probably indicating major intersections. A similar

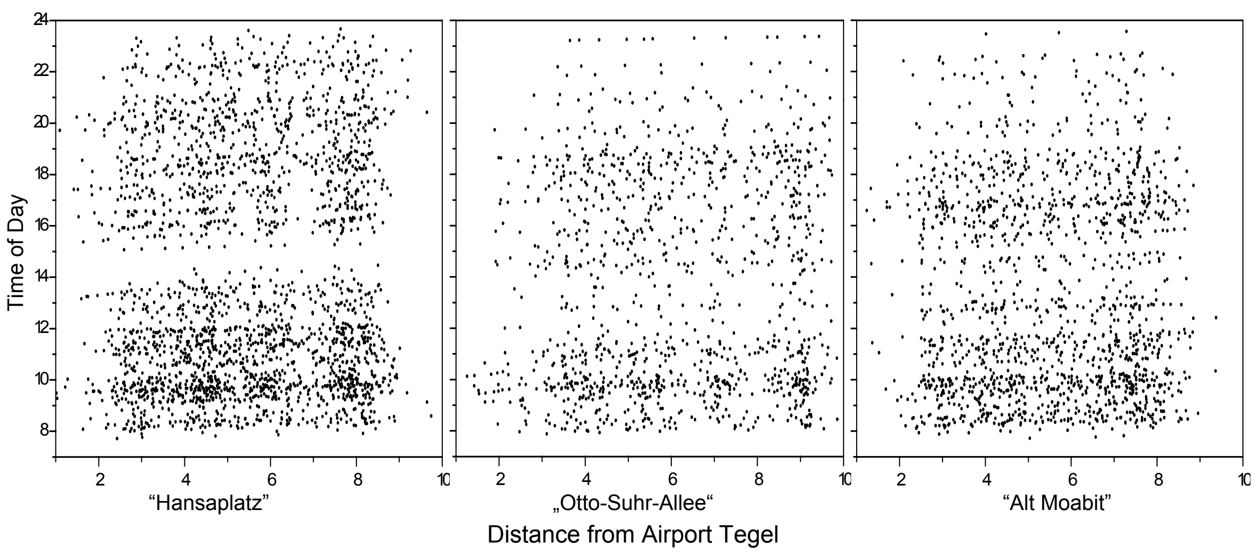

Figure 6: Measurement points of taxi positions in dependence of the taxis' distance from Airport Tegel. From left to right: "Otto-Suhr-Allee", "Hansaplatz", "Alt Moabit".

observation can be done for "Otto-Suhr-Allee", where especially around 10 am some clustering can be found. In comparison, "Alt Moabit" shows a more even distribution along the route, indicating a slow average speed or a large number of possible stops, respectively, or eventually a combination of both. Therefore, "Alt Moabit" may have the disadvantage of a sensed slow advance, which results in the preference of "Hansaplatz" even when the travel times on both routes are comparable.

5. Summary

The routes of taxi drivers have been investigated in order to get an insight into route choice behavior of car drivers which are tempted every day to find their best route. Taxi drivers are specialists in using the road network due to their high experience and their up-to-date knowledge of traffic problems. Therefore, they belong to a class of fairly informed drivers, consisting of experienced drivers and of drivers with up-to-date information on the network state, provided e.g. by broadcasting stations or other service providers.

Data from daily tours between the Airport Berlin-Tegel and the area next to Friedrichstraße have been investigated in terms of travel time, travel distance, average speed, taxi fee and parameters connected with the personal perception of the drivers, e.g. the "smoothness" of their routes. The investigations demonstrate, that the analysis of historical floating car data by statistical methods can provide useful information on route choice behavior, provided a high rate of GPS-equipped vehicles and a sufficient historical data base.

The results show expected as well as unexpected behavior of the drivers. At first, as expected, route choice behavior is primarily determined by the route length and the travel time needed. Clearly observable changes of the travel time during the day lead to a change of the preferable route. However, if routes are comparable in length and travel time an unexpected distribution is observable, where one of the routes is clearly preferred. Also, other routes with longer travel distance are not completely negligible. Whereas for the latter a comparable travel time is the most probable reason (and perhaps a higher fee), no objective distinction between the comparable routes is provided. The "smoothness" of the routes, a parameter introduced to describe the combination of number of stops, waiting time at stop lights and average speed between the stops, possibly shifts the decision towards a route, which exhibits sections with a higher average speed. Routes exhibiting an even low average speed are therefore more unfavorable.

Additional insight will be gained from a comparison of the route choices of the taxi drivers during the day with routes proposed by a router based on the historical data base. Also, the same origin-destination relation with unoccupied taxis, the inverse driving direction or other promising relations can provide more information. These approaches will be investigated in future, and can additionally be backed up with interviews of the drivers on their methods to determine the preferred route.

References

[1] Ralf-Peter Schäfer, Kai-Uwe Thiessenhusen, Peter Wagner: A traffic information system by means of real-time floating-car data, Proceedings of the ITS World Congress 2002, October 11-14, Chicago, USA

[2] Ralf-Peter Schäfer, Kai-Uwe Thiessenhusen, Elmar Brockfeld, Peter Wagner: Analysis of travel times and routes on urban roads by means of floating-car data, Proceedings of the European Transport Conference PTRC, September 4-9, Cambridge, UK