Deduction of passengers’ route choice from smart card data*

Evelien van der Hurk¹ and Leo Kroon¹ and Gábor Maróti² and Peter Vervest¹

Abstract—Deducing passengers’ route choice from smart card data provides public transport operators the opportunity to evaluate passenger service. Especially in case of disruptions when route choice models may not be valid this is an advantage. This paper proposes a method for deducing the chosen route of passengers based on smart card data and validates this method on a real life data set. The method reaches an accuracy of about 90 percent, also in case of disruptions. Moreover, it is shown how this method can be used to evaluate passenger service by a case study based on a real life data set of Netherlands Railways, the largest passenger railway operator of the Netherlands.

I. INTRODUCTION

Passenger service is an important factor in evaluating an operator’s performance. Passenger service defined in terms of travel time and vehicle crowdedness depends on passenger route choice. Most passenger route choice models are based on utility maximization [1] or regret minimization [2]. In case of sudden changes in the planning or disruptions in the public transport network, these models may not be valid. For one, passengers may not travel as predicted by the existing models due to the lack of up-to-date information. Secondly, the urgency to make quick decisions may result in unexpected travel paths. Thus, passenger service evaluated based on the traditional models may be incorrect. Therefore, a different method is required to analyze passenger route choice and thus passenger service in case of disruptions.

New data sources generated by automated fare collection systems allow for a data driven study of passenger route choice. These systems register the time and location of the start and end of each journey made by a smart card. In contrast to classical data collection methods such as travel diaries and surveys, these recently introduced systems store all passengers’ journeys every day. Moreover, data resulting from conductor checks generate additional data on the route choice of a passenger by registration of checked cards including time and train number.

This paper contributes to smart card data research [3] in proposing a method for route deduction, that deduces the chosen route of passengers from smart card data, and validating this method through additional data resulting from conductor checks. It extends previous research on route deduction using smart card data [4] by including a validation of our proposed methodology and by studying passenger service in disrupted situations based on this method.

Our proposed methodology has a performance of over 90 percent, indicating a correct match between passenger and train, in our case study based on a real life data set of Netherlands Railways (NS). We find that the route deduction works well based on the planned timetable and the realized operations, as well as in case of disruptions. Moreover, we show how passenger service, based on passenger route choice, can be analyzed based on our route deduction method. This provides valuable managerial insight into the consequences of changes in planning and the effects of disruptions.

II. PROBLEM DESCRIPTION

Route deduction, coupling journeys registered in the smart card data to a specific route in the timetable, is complicated. For one, because passengers are known to not always choose the shortest route. Secondly, multiple routes often fit within the registered time of entrance and exit, as passengers may undertake different activities than traveling in this time. Thus the route deduction requires both a method for finding optional routes and a method for selecting the correct route in case multiple options remain for a journey in the specified time interval.

This paper addresses both the route finding as well as the coupling of journeys to routes in case both a start and end registration of time and location of the journey is available. A journey is a passenger traveling from origin to destination during a specific time interval. A route is a set of specific trains leading from origin to destination with an associated departure and arrival time.

Evaluation of the method for route deduction is based on a 5 day NS data set consisting of about 500,000 journeys with a conductor check. The analysis of passenger behavior is based on a larger data set containing all journeys made by smart card on these days. The 5 days were chosen to contain two regular days, and three special days. On the latter, the schedule was changed due to expected bad weather conditions. These changes were announced to the passengers at 17:00h on the previous day.

A changed schedule contains approximately 20 percent less trains and reduces the frequency of trains by half on the main corridors. The schedule is changed in order to prevent major disruptions and add-on effects due to the weather. Though it is clear that this change in operations has increased performance since its introduction two years ago, still the punctuality of trains on these days is significantly below average. Therefore it provides a good case study for

*This work was supported by The Netherlands Science Foundation NWO
¹E. van der Hurk, L. G. Kroon and P. H. M. Vervest are with Rotterdam School of Management, Department of Decision and Information Sciences, Erasmus University, 3000 DR Rotterdam, The Netherlands {ehurk, lkroon, pvervest} at rsm.nl
²G. Maróti is with the Netherlands Railways, gabor.maroti at ns.nl

©2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
testing the performance of the route deduction and analyzing passenger service in case of disruptions.

III. DATA

This section describes the used data: smart card data (Section III-A) contains the information on passenger journeys, timetable information (Section III-B) defines the available routes, and data from conductor checks (Section III-C) are used for validation.

A. Data on passenger route choice

NS’s smart card system requires passengers to tap their card to a machine positioned at the station at the beginning and end of each journey, referred to as checking-in and checking-out. This system is similar to such systems implemented for metro, train and subway in respectively London, Tokyo, and Seoul. A journey \( j \) contains a check-in time \( t_{ij}^c \), a check-in location \( s_{ij}^c \), a check-out time \( t_{ij}^a \) and a check-out location \( s_{ij}^a \). The analysis focusses on complete journeys that on average account for 95% of all transactions, on special days the change in distribution was less than 3%. Table I presents an example of smart card data. All journeys can be linked to a card \( c \). This link is used to check the route deduction based on the conductor check data, described in Section III-B.

<table>
<thead>
<tr>
<th>Date</th>
<th>Card</th>
<th>Time</th>
<th>Station</th>
<th>Check-in</th>
<th>Time</th>
<th>Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 Jan</td>
<td>1234</td>
<td>7:55</td>
<td>A</td>
<td></td>
<td>8:35</td>
<td>B</td>
</tr>
<tr>
<td>10 Jan</td>
<td>1234</td>
<td>18:05</td>
<td>B</td>
<td></td>
<td>18:30</td>
<td>A</td>
</tr>
<tr>
<td>10 Jan</td>
<td>2345</td>
<td>12:07</td>
<td>E</td>
<td></td>
<td>12:20</td>
<td>F</td>
</tr>
</tbody>
</table>

Smart card data stores information on both location and time of a journey for all journeys made by such a card, for every passenger that uses the card, every day. In the near future the smart card is expected to become the only method of payment. Currently usage of smart card depends on the product type, specifically regular travelers are less represented in the data. During time of analysis, about a third of the passengers’ journeys are accounted for in the smart card data. Neither seat reservation nor reservation for a specific train is available in the Dutch Railway System, thus smart card data is the only source of time-dependent journey information.

B. Timetable information

Timetables contain information on the departure and arrival times of trains at stations, a sample of which is presented in Table II. Passengers plan their journey based on the planned timetable. The realized timetable differs from the planning due to disruptions and disturbances. Therefore the actual route of a passenger depends both on the planned timetable and the realized timetable. In this research, we therefore compare route generation based on the realized and the planned timetable.

<table>
<thead>
<tr>
<th>Date</th>
<th>Train number</th>
<th>Departure Time</th>
<th>Departure Station</th>
<th>Arrival Time</th>
<th>Arrival Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 Jan</td>
<td>405</td>
<td>7:55</td>
<td>A</td>
<td>8:05</td>
<td>B</td>
</tr>
<tr>
<td>10 Jan</td>
<td>405</td>
<td>8:05</td>
<td>B</td>
<td>8:20</td>
<td>C</td>
</tr>
<tr>
<td>10 Jan</td>
<td>708</td>
<td>12:07</td>
<td>E</td>
<td>12:20</td>
<td>E</td>
</tr>
</tbody>
</table>

C. Conductor checks

NS employs conductors who, as part of their duty, check the validity of tickets of passengers. The validity of a smart card ticket is checked using a Mobile Chip card Reader (MCL). It states whether a card has checked in, thereby validating the ticket. The device then stores the card-id, the time, and the train number of the current train of the conductor. MCL-data, as presented in Table III, thus provides information on the route choice of the passenger as it registers a card-id together with a date, time and train number. MCL-entries are defined as a single check for a card, corresponding to a row in Table III.

The routes of the passengers cannot directly be deduced from the MCL-data. This is because for one the tickets of only a subsample of passengers are checked. Secondly, in some cases multiple routes may still exist given an MCL-entry and a journey in case of transfers. However, MCL-data can be used to validate the route deduction of journeys to routes.

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Train number</th>
<th>Card number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Jan 2013</td>
<td>7:55</td>
<td>501</td>
<td>123456789</td>
</tr>
<tr>
<td>1 Jan 2013</td>
<td>8:05</td>
<td>408</td>
<td>987654321</td>
</tr>
<tr>
<td>2 Feb 2013</td>
<td>12:07</td>
<td>705</td>
<td>123456789</td>
</tr>
</tbody>
</table>

IV. METHOD

The route deduction and validation process consists of the three steps presented in Fig. 1. In the first step a set of time-and-date-specific routes per OD is generated based on the timetable, this can be either the planned or the realized timetable. Secondly, for each journey \( j \) a route matching the time interval is selected from this global route set, we test different selection rules. Finally, we evaluate the performance of this method using the MCL-data as a ‘ground-truth’ data set.
A. Step 1: route generation

A trip is a train ride between two consecutive stations; the timetable defines the list of all trips. A route is an ordered set of trips along which a passenger can travel from an origin station to a destination station at a specific time.

The timetable is translated into an event-activity-network $\mathcal{N} := (\mathcal{E}, \mathcal{A}), \mathcal{E} := \mathcal{E}_{dep} \cup \mathcal{E}_{arr}, \mathcal{A} := \mathcal{A}_{drive} \cup \mathcal{A}_{wait}$, first introduced by [5]. Events correspond to time dependent departures and arrivals of trips. Arcs represent the activities of a train driving to the next station or passengers waiting at a station. In this event activity network, stations are represented as several events over time, namely when trains arrive and depart from this station, with waiting arcs in between. A route corresponds to a directed path in $\mathcal{N}$ leading from the origin station to the destination station.

Passengers are known to choose a route based on multiple criteria, such as the travel time and the number of transfers. This choice set is unknown to the operator and as this set consists of more than the shortest path, generating such a set is a complex problem [6]. Finding shortest paths in a graph is however a relatively easy task. Therefore, for each origin-destination pair we generate a set of time dependent routes $\mathcal{R}^{OD}$ for the full day by:

- selecting all trips (activities) departing from the origin station on one specific day
- finding for each of these trips the shortest path starting with this trip and leading to the destination without returning to the origin station

This results in a diverse set of routes in space and time. Some of these may be less likely as they contain a large detour. Thus we reduce this set by eliminating all routes where there exists another route in the set $\mathcal{R}^{OD}$ that in comparison arrives no later, departs no earlier, and has not more than one additional transfer. More conservative rules we tested, extending the set of routes, performed worse than this rule.

The MCL data can be used as ground truth to test both the route selection and the route generation. The route generation is validated by checking if there exists a route in $\mathcal{R}^{OD}$ that matches the MCL-check. If there is not such a route, but there exists such a route in the graph $\mathcal{N}$, we know the route generation is incomplete. During this check, we can simply extend the route set with the valid route in the graph. Though this is not a long term valid adjustment to the method as it creates dependency between the validation and the route deduction method, this allows us to judge whether improving the route generation by e.g. using a learning algorithm will improve the route deduction. As the timetable changes slightly from day to day, using past data to generate routes is not straightforward and will be part of future research.

We restrict the route generation and coupling to those OD pairs with on average at least three journeys a day. This accounts for 98 percent of all journeys, but reduces the number of OD pairs to less than a tenth. Though the approach can easily be extended to the full set, this restriction saves computation time (13 minutes for the limited set versus approximately 5.2 hours for the full set per day) and prevents memory issues on the local server session (Citrix) we were restricted to using due to the sensitivity of the data.

B. Step 2: route selection

For each journey $j$ in the smart card data we select a group of candidate routes $\mathcal{R}^j$ from the set of OD routes $\mathcal{R}^{OD}$ that fit within the registered check-in and check-out time.

Over 50 percent of the journeys admit multiple candidate routes. In order to find a unique one, we consider four route selection rules:

1) First Departure ($FD$)
2) Last Arrival ($LA$)
3) Least Transfers ($LT$)
4) Selected Least Transfers Last Arrival ($STA$)

$FD$ selects the route that departs closest to the time of check-in. This route minimizes the waiting time at the departure station. $LA$ selects the route arriving closest to the time of check-out. This route minimizes the waiting time at the arrival station. $LT$ selects the route with the smallest number of transfers. If any of the rules $FD$, $LA$ and $LT$ finds multiple equally good routes then we use $LT$, $LA$ and $FD$ (in this order) to break the tie. The order based on performance of the separate rules was initially defined based on company knowledge, and later confirmed by our computational results.

Finally, $STA$ is somewhat different from the other methods. It uses the hypothesis that most passengers spend their time traveling between check-in and check-out, therefore quicker alternatives might be less likely when a slower route fits well within the check-in and check-out. But, passengers still prefer direct trains and tend to check-out closer to arrival than to departure. Thus, simply maximizing the travel time will not always lead to the best route and may select a route with unnecessary transfers. STA first selects routes where the sum of difference between the check-out and the route’s arrival, together with the check-in and the route’s departure, is less than 10 minutes. The value of 10 minutes is motivated by the fact that the main corridors have 4 trains per hour. When
multiple routes remain, we use $LT$, $LA$ and $FD$ (in this order) to break the tie. If the set after selection is empty, all routes fitting within the check-in and check-out time are selected and we use again $LT$, $LA$ and $FD$ (in this order) to break the tie.

C. Step 3: validation

The route deduction is validated on the subsample of journeys that have a registered belonging MCL-entry. A route matches the MCL-entry when the route contains the train number at the time registered in the MCL-data. Multiple checks for one journey are scored separately. In case multiple routes match the MCL-entry, the performance is still determined by only the linked route to the journey and the MCL-entry.

We define the performance of the route deduction as the percentage of journeys with a matching route to the MCL-entry, in comparison to the total set of journeys with an associated MCL-entry. We measure performance based on the 95 percent of these journeys that have at least one and no more than five possible routes to choose from in the route selection. Some journeys have over a hundred routes associated to it due to registration errors or because passengers forget to check out. Limiting the set in this way prevents this to influence the evaluation of the route deduction method.

In case the route does not match an MCL-entry, we distinguish four cases:

1) A matching route exists in the set $R_j$, but the method chooses incorrectly.
2) A matching route exists in the route set for the OD pair $R^{OD}$, but does not fit within the registered check-in and check-out time of the journey.
3) A matching route does not exist in the set $R^{OD}$, but does exist in the timetable.
4) A matching route that fits within the registered entrance and exit times of the journey does not exist in the timetable.

In theory there should always be a route that has the passenger in the train at the time the ticket was checked by the conductor, and therefore cases 2 and 4 could not occur. However, in our data set they do occur. Upon closer inspection it may be verified that this is due to a dissynchronisation of clocks in the respective data sets.

For over 10 percent of the smart card journeys there does not exist a route in the timetable that connects the origin and destination of the journey within the registered entrance and exit time. As the timetable is considered to be the most reliable source, a global correction was applied to the smart card data, extending the check-out time with three minutes. After this extension, case 4 occurs for less than 1.5% of the journeys. A global check-in time correction did not improve results. This is likely due to the fact that for some trains and platforms the check-in device is very close to the location of the departing train. As this takes only seconds and we define time in minutes, a global correction did not improve results.

Synchronizing the MCL-data to the timetable is more difficult because it is not known what causes the dissynchronization. Therefore we leave this to future research, and accept a loss of around 5 percent of MCL-data due to registration errors.

We measure the performance based on the set of journeys that have a matching route or an error corresponding to case 1 or case 3. Case 1 results from an error in the selection rule, discussed in Section IV-B. Therefore we compare the performance of different selection rules. Case 3 is caused by an incomplete route generation in Step 2, discussed in Section IV-B. Therefore we compare the performance of route-generation based on the timetable graph and the extension of this route-set based on MCL data. The second method keeps the selection of a route for a journey still independent of the MCL-entries. It is therefore an indication of the benefit of improving the route generation method, and can in its current form be used directly for any future day of smart card data.

V. RESULTS

The method for route deduction and validation is tested for a 5 day sample of real life data of NS, containing two regular days and three special days with reduced service due to expected extreme weather condition. Changes to the timetable were communicated to the passengers at 17:00h the previous day, and despite the change these days had a below average punctuality. We compare performance of different selection rules based on realized and planned timetable, and for the two route generation methodologies (Section IV-A).

Passenger service and difference in passenger behavior on the special days is studied based on the route deduction. Results show that passengers adjust their departure time to the changed timetable over time, although still the major decrease in service results from a longer waiting time at the station. A detailed discussion is presented in Section V-B.

A. Linking journeys to a route

The performance of the route deduction depends on two aspects: one, the generation of all potential routes followed by the passengers, and second, the choice of the correct matching route. We find that the route generation method, generating around 400,000 routes, contains most but not all routes. The extended list, where routes for journeys with case 3 error were added to the original set, increases performance of the coupling. Adding these 8000 routes (2 percent of the original set size) leads to a performance increase of up to 6 percent of the coupling based on this list. Table IV presents results for the created and extended list, based on the planned and realized timetable, for all rules on one regular day. We find that the performance of the route deduction improves using this extended set, and is slightly better when using the realized timetable. Note that the addition of more routes does not necessarily lead to a better performance, as it increases the choice set per journey. However, the results of the extended list indicate that the route generation influences the performance of the route deduction.
TABLE IV

PERFORMANCE OF ROUTE DEDUCTION IN PERCENTAGE FOR THE CREATED AND EXTENDED ROUTE LIST, BASED BOTH ON THE PLANNED AND REALIZED TIMETABLE, ON REGULAR DAY 1 FOR ALL RULES.

<table>
<thead>
<tr>
<th>Method</th>
<th>Planned Timetable</th>
<th>Realized Timetable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Generated List</td>
<td>Extended List</td>
</tr>
<tr>
<td></td>
<td>Generated List</td>
<td>Extended List</td>
</tr>
<tr>
<td>FD</td>
<td>77.8%</td>
<td>75.5%</td>
</tr>
<tr>
<td>LA</td>
<td>77.8%</td>
<td>78.5%</td>
</tr>
<tr>
<td>LT</td>
<td>80.8%</td>
<td>82.1%</td>
</tr>
<tr>
<td>STA</td>
<td>83.8%</td>
<td>84.6%</td>
</tr>
</tbody>
</table>

The rules STA and LT, preferring direct connections, outperform the other rules LA and FD. Results in Table V show that STA has the best performance, even though for special days the overall performance is slightly less than on regular days. The relatively good performance on special days can be explained from the fact that due to the reduction in train frequency, the choice set per journey is smaller, and hence the probability to choose the correct route is higher.

The presented results are based on the realized timetable and extended list, which are the best settings. Results are similar for other settings. Thus we can conclude that STA is a successful method for coupling a journey to a route, also in case of special events.

To analyze whether the accuracy depends on the OD pair, we assigned a performance score for each OD pair. This score is weighed by the number of journeys belonging to the OD pair. We found that less than two percent of all journeys have an OD performance below 50 percent. Moreover, almost a third of journeys have an OD with a perfect score of 100 percent, while over two thirds of the journeys have an OD performance of 90 percent or more. Note that as these scores are calculated per OD, they cannot directly be compared to the number of options per journey. Fig. 2 shows the cumulative performance score per OD on the horizontal axis, and the percentage of journeys with this minimal score on the vertical axes. The overall good performance is indicated by the slow decline of the line. The perfect score for over 30 percent of the journeys is indicated by the end of the graph at (1, 0.33), indicating that 33% of the journeys has an OD with a performance of 100 percent.

B. Analyzing change in passenger service

Based on the route-deduction, we can now compare on-route time differences on the special days with the journey-time differences of the special days, thus providing a deeper insight into the experienced passenger service. We focus on the following two aspects:

- a change in the duration of the route, due to disruptions, disturbances and additional transfers
- a change in the duration of the journey, due to increased waiting time at the stations

The first is referred to as route-time, the difference between the arrival and departure times of the route assigned to a journey based on the route deduction. The second is referred to as journey-time, the difference between check-in time and check-out time of a journey, which actually contains both the additional route-time as well as the additional waiting time at the station. It is important to distinguish the route-time from the journey-time, as it is the difference between provided logistic-service and the experienced service of the passenger. Especially with a reduction of the frequencies of trains this distinction is important.

We evaluate the passengers’ service as the difference between normal and special days, by calculating the difference in journey-time and the difference in route-time for these two types of days. Differences are calculated per OD-pair, to prevent a change in the OD matrix to cause a change between special and normal days. Differences per OD pair are weighed by the number of journeys on the special days, to obtain the average change in service for all journeys on the special days. Finally, differences are evaluated based on departure times of journeys. By calculating the difference in route-time and journey-time based on 15 minute departure-time intervals, the difference in passenger service in peak and off-peak hours can be compared.

We find that indeed special days have both a longer journey-time and a longer route-time in comparison to regular days. The two regular days are not significantly different in this respect. Fig. 3 contains graphs of the average difference in journey-time (top) and route-time (bottom) for each of the three special days. Further analysis based on the route-deduction shows that the longer journey time mostly...
results form a longer waiting time at the departure station, while the waiting time at the arrival station is statistically the same for all days. Moreover, it seems that passengers learn over time to adjust their departure time to this new schedule, while at the same time punctuality increased a bit, both leading to better passenger service on later days.

We compared different route selection rules for the route deduction. We found the route deduction to perform at an accuracy of over 90 percent for the best selection rule, STA. The method performs well on both regular and normal days. Comparing the route generation on the planned timetable with the realized timetable, we find that the realized timetable performs slightly better. The route generation seems to contribute significantly to the performance of the timetable. Adding missing routes based on the MCL-data increases the performance. Therefore, improving the route selection algorithm may improve the route deduction. Future research will focus on developing a learning algorithm for the route generation module based on the historic MCL-data base. Moreover, it will study whether an individual coupling algorithm, using the data on previous route choices per card, can further improve the route deduction.

The route deduction is sensitive to the synchronization of the clocks of the different data collection systems of smart card data, timetables and MCL-entries. We find that a global correction of the clock in the smart card data helps to synchronize the timetable with the smart card data. Further research may be able to synchronize the MCL-data with the timetable as well, thereby extending the test set and improving the performance of the validation.

Passenger service, resulting from the interaction between passengers traveling and the operations, is analyzed based on route deduction. Studying the change in passenger service in terms of a change in route-time and journey-time we find that these are both higher in case of special days. Specifically, passengers are waiting longer at the station before their route-departure time. The comparison of the special days suggests that passengers adjust over time to this changed schedule. The presented approach can also be applied more broadly to evaluate change in passenger service either distributed over the day or per day.

VI. CONCLUSIONS AND DISCUSSION

This paper proposes a method for route deduction from smart card data. The route deduction is important for analyzing passenger service level, which is dependent on route choice. Though smart card systems have been implemented around the world, not much research has been carried out on deriving route choices from this data. Moreover, thanks to the data from conductor checks, the route deduction is validated on a card specific level. The card specific validation is thus of a higher accuracy than previously suggested validation methods based on vehicle loads [4], especially when not all passengers in the vehicle are traveling by smart card. To our knowledge, this is only the second paper apart from [4] to propose a method for route deduction, and it is the first paper to include validation of this method, to focus on how disruptions and the usage of different timetables affect the route deduction, and to use this route deduction to evaluate passenger service in case of disruptions.

ACKNOWLEDGMENT

We thank Netherlands Railways for their collaboration and the provision of data for this research.

REFERENCES