






Identifying Relevant Transfer-Connections from Entry-Only Automatic Fare Collection Data: The Case Study of Porto

Joana Hora^{1,2}(✉) , Teresa Galvão^{1,2} , and Ana Camanho^{1,2} 

¹ Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

² INESC TEC, Porto, Portugal

joana.hora@gmail.com,
{acamanho, tgalvao}@fe.up.pt

Abstract. The synchronization of Public Transportation (PT) systems usually considers a simplified network to optimize the flows of passengers at the principal axes of the network. This work aims to identify the most relevant transfer-connections in a PT network. This goal is pursued with the development of a methodology to identify relevant transfer-connections from entry-only Automatic Fare Collection (AFC) data. The methodology has three main steps: the implementation of the Trip-Chaining-Method (TCM) to estimate the alighting stops of each AFC record, the identification of transfers, and finally, the selection of relevant transfer-connections. The adequacy of the methodology was demonstrated with its implementation to the case study of Porto. This methodology can also be applied to PT systems using entry-exit AFC data, and in that case, the TCM would not be required.

Keywords: Public Transportation · Transfers · Automatic Fare Collection

1 Introduction

The decisions made at the Transit Planning Process (TPP) are grounded on passengers' behavior assumptions, such as the expected demand of passengers. These assumptions are mainly drawn from the analysis of historical records or surveys. In its turn, the implementation of TPP decisions impacts the Public Transportation (PT) service delivered to passengers, e.g., with changes in routes' design, frequencies or schedules. Finally, changes in the PT service will impact the behavior of passengers, which is not deterministic and often does not evolve as expected (e.g., choosing to commute with private car or PT, or choosing between alternative PT routes when several options are available for the same Origin-Destination (OD)). Figure 1 shows this causal cycle interrelating the TPP, the PT service, and the behavior of passengers.

Several sequential stages integrate the TPP. The Network Design (ND) stage returns the set of routes composing the PT network, designed to provide the best possible transportation service by meeting the passenger demand. The Frequencies Setting (FS) stage assigns frequencies for all daily Uniform Demand Period (UDP) on each route (e.g., assign one vehicle every 15 min during a morning peak). The Timetabling (TT) stage returns the timetables with the departure and arrival times of all daily trips on each route, typically at the level of Time Control Point (TCP)s or stops. Follows the Vehicle Scheduling (VS), the Driver Scheduling (DS) and the Driver Rostering (DR) stages. Although aspects such as passenger demand can never really be known or accurately described by its historical observations, the adoption of assumptions related to them is needed to sustain decision-making at any TPP stage.

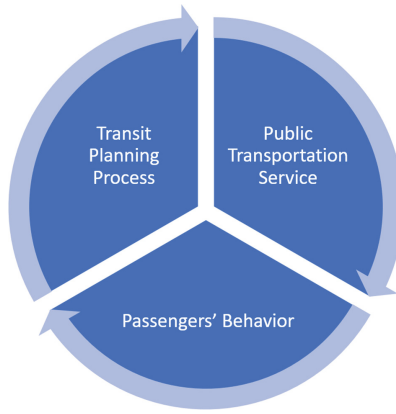


Fig. 1. Cycle ‘Transit Planning Process - PT service - Passengers behavior’.

The technological advent of the last decades had endowed planners and decision-makers with access to a higher volume of accurate data, such as Automated Fare Collection (AFC) records, fostering its application into research and development activities. At the same time, improved computational tools have increasingly been applied to solve TPP problems.

There are different techniques used for TT, depending on the experience and resources of planners and companies [1]. One popular approach is the implementation of the Synchronization Timetabling Problem (STP) [2–4]. The STP builds timetables pursuing the reduction of the overall inconvenience for passengers. The idea is to obtain coordinated timetables that enable smooth interchanges through the minimization of passengers’ waiting-time and bunching of vehicles.

The STP is usually applied to simplified, yet realistic networks. This simplification is considered not only due to the complexity of the STP (i.e., NP-hard [2]) but also because increasing the size of the network significantly reduces the flexibility of the solutions obtained, which is critical for finding compatible solutions in the TPP downstream stages, especially at the VS.

This work aims to identify the most relevant transfer-connections within a PT network, which can be further used to build a *simplified yet realistic* representation of the routes that should be coordinated to provide a quality service to passengers.

A methodology is proposed to identify relevant transfer-connections from entry-only AFC records. The methodology has three main steps: the implementation of the Trip-Chaining Method (TCM) to estimate the alighting stops of each AFC record, the identification of transfers, and finally, the selection of relevant transfer-connections. Relevant transfer-connections are selected considering four main assumptions: (1) identification by experts, (2) in case of shared paths, favor the selection of connections positioned at strategic stops such as merging or crossing routes, (3) compliance with a maximum walkable distance threshold, and (4) compliance with a specified threshold of demand.

The TCM estimates the alighting stops of entry-only AFC records. It considers the sequence of trips made by each passenger in each day, connecting trip-legs of each smart-card. The literature on TCM counts with several implementations at different PT systems worldwide, differing mainly in the set of assumptions implemented [5–9]. The majority of these works keep the two grounding assumptions proposed in the seminal work of [5]: (1) most passengers will start the next trip of the day at or near the alighting stop of their previous trip, and (2) most passengers end the last trip of the day at or near the boarding stop of their first trip of the day.

The identification of transfers has also been addressed in literature considering assumptions of transfer walking distance [6, 8, 10], transfer time thresholds [8, 10, 11] and transfer network feasibility conditions [10].

2 Concepts: Transfer-Node, Transfer-Connection and Transfer-Event

This work distinguishes the concepts of transfer-connection and transfer-node. A transfer-node is the geographic area where two or more routes meet, cross, or merge. A transfer-connection refers to the possible interchange of passengers between two specific directed-routes, possibly separated by a walkable path. A transfer-event is the observation of passengers transferring through a transfer-connection, with specific detail on the vehicles involved and on time.

2.1 Transfer-Event

Figure 2 schematizes a transfer-event. A transfer-event has four main moments: (1) a Feeding Vehicle (FV) from the Feeding-route (FR) arrives and passengers alighting; (2) passengers walk between the alighting-stop and the boarding-stop when a walkable path exists; (3) passengers wait at the boarding-stop; (4) a Receiving Vehicle (RV) from the Receiving-route (RR) arrives and passengers board.

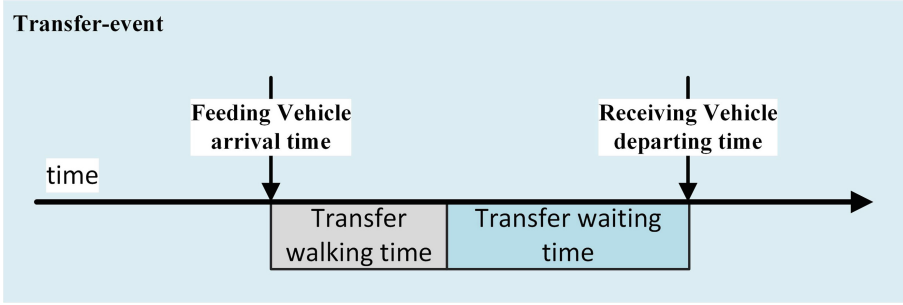


Fig. 2. Scheme of a transfer-event.

Several runs serve each directed-route over a day. A transfer-event addresses the interchanging process between two specific trips, operated by specific runs of each directed-route. A transfer-event encompasses knowledge on the specific time for the FV arrival and the RV departure.

In the case when the FV arrives earlier than scheduled and the RV is on-time, passengers are unlikely to miss the RV. Passengers willing to perform this transfer-event will have extra transfer waiting-time, or in some cases, they might be able to board a prior run of that receiving-directed-route.

In the case when the FV arrives later than scheduled and the RV is on-time, passengers are likely to lose the transfer-event and wait for the next run of that receiving-directed-route, or they might board the RV with almost zero transfer waiting-time.

Many other transfer-event scenarios can be studied considering different FV and RV arrival and departing-time. The study of transfers is of utmost importance to enhance as much as possible successful transfer-events, reducing transfer waiting-time that is inconvenient for passengers, and improve overall passenger flow within the PT network.

2.2 Transfer-Connections and Transfer-Nodes

The simplest case of a transfer-node is the case where two route-terminus meet, as illustrated in Fig. 3. The last stop of one route is at the same geographic area of the first stop of another route. This type of transfer-node is commonly found in peripheral areas of cities, aiming to connect PT service from the suburbs to strategic PT routes traveling into cities. In this particular case, the transfer-node encompasses only two possible transfer-connections, as identified in Fig. 3.

Another common type of transfer-node occurs when two routes cross or merge, as illustrated in Fig. 4. In both situations, the resulting transfer-node always includes eight transfer-connections, as identified in Fig. 4. This analysis deliberately excludes any interchanging of passengers between trips of the same route, regardless of route direction. The main reasoning is that passengers would only board into the same route at a consecutive trip in case of (i) a mistake, or

(ii) the start of a new journey which takes place after an activity (even when the two consecutive trips occur within a short duration). Either way, they do not reflect an interchange and therefore are not included in this analysis.

Figure 5 represents the case of two routes sharing a segment path, with their merging and splitting transfer-nodes. Possible transfer-connections at the merging/splitting transfer-nodes are represented at boxes (i) and (iii), while possible transfer-connections at the shared path are represented at the box (ii).

Considering the case of Fig. 5, although passengers can interchange over the shared path represented in box (ii), the methodology followed in this work considers that the transfer-connections positioned at the merging and splitting transfer-nodes (boxes i and iii) should be given priority with respect to any stop positioned at the shared path (box ii). This concept is further included as an assumption to identify relevant transfer-connections. The overall goal is to optimize passenger waiting-time and vehicle congestion, which is achieved more efficiently by concentrating transfers in a reduced number of strategic stops.

Finally, the analysis of transfer-connections when three or more routes intersect or merge at the same geographic area is easily understood with a matrix approach, as the one exemplified in Table 1. For example, a transfer-node crossed by three routes has 24 possible transfer-connections, and a transfer-node crossed by four routes has 48 possible transfer-connections.

3 Methodology to Identify Relevant Transfer-Links from AFC Data

This section details the methodology adopted to identify relevant transfer-links from entry-only AFC data. This methodology embodies three main steps: (1) implementation of the TCM to estimate alighting stops for all AFC records, (2) the subsequent application of criteria to identify transfers, which also allows to link trip-legs and reveal OD, (3) the identification of transfer-links of improved relevance regarding further consideration for further optimization techniques, particularly the synchronization.

The data-set of AFC records is sorted by smart-card Unique Identifier (UID) and then chronologically. The following two steps consider AFC records sequentially in this order. These steps are schematized in Fig. 6, and detailed in Sects. 3.1 and 3.2.

The third step is performed independently of the first two steps. After the estimation of alighting stops and transfer-connections for all AFC records, the identification of relevant transfers will be carried out in a new algorithmic procedure detailed in Sect. 3.3. Figure 7 schematizes this procedure.

3.1 TCM Implementation

The TCM allows to estimate the alighting stops for each AFC record. The TCM implementation adopted in this work follows all details provided in [12]. The main assumptions adopted for this implementation are detailed in Table 2.

When there are two or more AFC records for the same smart-card, the algorithm proceeds to estimate their alighting locations, applying the TCM. When there is only one AFC record, the TCM cannot be applied, and the algorithm cannot estimate the Destination of that trip. In that case, the algorithm proceeds to the next smart-card UID in the data-set.

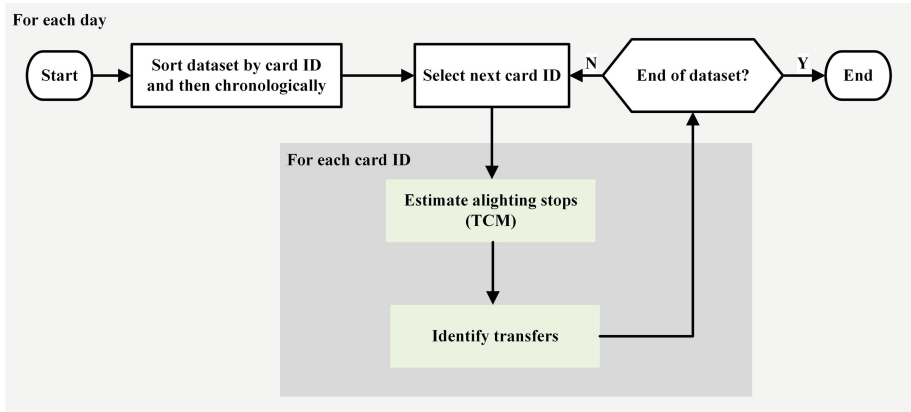


Fig. 6. Methodology followed to estimate alighting stops and identify transfers from entry-only AFC data.

When there are two or more AFC records for the same smart-card, the algorithm continues by selecting the next record. Knowing the boarding stop, route, and direction, the algorithm specifies as possible alighting stops all of the stops that were not yet traveled in that path. If the current AFC record is not the last of the day, the TCM estimates that the passenger alights, from the set of possible alighting, at the stop with the shortest walking distance regarding the boarding stop of the subsequent record. If the AFC record is the last of the day, the same approach is adopted but regarding the boarding stop of the first AFC record of the day (assuming the passenger would travel back home at the end of the day).

When the walking distance between the current alighting stop and the subsequent boarding stop is higher than the threshold of 3 km, we assume that the passenger traveled off the transportation system between these two AFC records. For example, picking up a ride or using another transportation mode. In that case, the estimation made regarding its alighting location is discarded. Similarly, for the case of the last AFC record of the day, if the walking distance between the current alighting stop and the first boarding stop of the day is higher than 3 km, we assume that the passenger traveled off the transportation system on its return home and that estimation is discarded as well.

Table 2. Assumptions adopted to implement the TCM using entry-only AFC data.

	Assumption
1	Passengers start the next journey stage at or near the alighting location of their previous trip
2	Passengers end the last trip of the day at the boarding location of the first trip of the day
3	Passengers can only alight in the sequence of stops not yet traveled by the route direction they boarded
4	Passengers travel off the transportation system when the walking distance between consecutive AFC records is higher than a specified threshold (in km)

3.2 Identify Transfers

In this work, we aim to identify which trip-legs are linked by transfers, therefore identifying real Origins and Destinations incurred by passengers. Therefore, the algorithm proceeds by distinguishing if the alighting stop of a AFC records corresponds to a transfer within a sequence of trip-legs, or if it corresponds to the Destination of a trip. The main assumptions adopted to identify transfers from AFC records, regarding the behavior adopted by passengers in their daily travel patterns, are detailed next.

Table 3. Assumptions to distinguish transfer-events from trip-ends.

	Assumption
1	Passengers will not transfer to another vehicle of the same route in which they are traveling, regardless of its direction
2	Passengers are not willing to walk more than a specified threshold to transfer to another route (in meters)
3	Passengers are not willing to wait for more than a specified threshold to transfer to another route (in minutes)
4	The boarding stop of the first AFC record of the day is the Origin of a trip
5	The alighting stop of the last AFC record of the day is the Destination of a trip
6	When passengers travel out of the system, the next AFC record is the beginning of a trip

Assumption 1 implies that passengers will only perform two consecutive AFC records on the same route when executing two different trips. That is, a passenger does not perform a transfer to board the same route he was already traveling, even if in the opposite direction, unless by mistake. This way, if a passenger

boards the same route in the consecutive AFC record traveling in the opposite direction, we consider that the passenger is performing a new trip and not a transfer. For example: (a) a passenger travels from home to bank, and then from bank to home using the same route, in less than 30 min; (b) a passenger travels from home to school, pick up the kids, and go back home using the same route, in less than 30 min. The main reasoning of this assumption is that it helps to distinguish transfers from trip ends. In these examples, there was an activity that took less than 30 min, and the walking distance between the two AFC records is lower than 200 m. Assumptions 2 and 3 are aligned with the literature on the topic.

Assumptions 4, 5, and 6 establish basic rules that identify trips' start and end. Assumption 4 considers that the boarding station of the first daily trip of each smart-card will always be the beginning of a trip. Analogously, Assumption 5 states that the landing station of the last daily trip of each smart-card will be a Destination of a trip. Finally, assumption 6 addresses situations in which passengers travel by alternatives to the PT system (e.g., by private cars or bicycles). When the estimation of the landing stop of a AFC record is discarded (the passenger traveled off the transport system), the following AFC record is always considered as the beginning of a new trip.

From the successful estimations of alighting stops other than the last trip of the day, the algorithm will distinguish between transfers and trip ends. To perform this distinction, we implemented three criteria, aligned with the assumptions previously defined.

For each pair of consecutive AFC records, the algorithm will assess: (1) if both records are from the same route, in that case, both records are considered to belong to different journeys. (2) if the walking distance is within the specified threshold; (3) if the time elapsed is within the specified threshold; The algorithm identifies a transfer when all three criteria are met. If at least one of these criteria do not meet, the first AFC record of the pair classifies as a trip end (its alighting stop is the trip Destination), and the boarding stop of the next AFC record is the Origin of a new trip.

3.3 Selecting Relevant Transfer-Connection

The methodology for the identification of relevant transfer-connections is grounded on four main assumptions. These assumptions are not perceived as rigid criteria, but instead as a framework to support the selection process. A description of the four assumptions is provided in Table 4.

Assumption 1 considers that the expertise and knowledge of distinguished stakeholders must be accounted in to identify relevant connections, even without meeting any quantitative criteria. This includes cases such as providing transportation service in areas with lower population density, maintain a transfer-connection that has existed for a long time and therefore is awaited by passengers, or to ensure the connection between the last trips of specific routes (allowing passengers that travel late to reach home).

Table 4. Assumptions for the selection of relevant transfer-connections.

	Assumptions
1	A transfer-connection is relevant when identified as such by experts, considering social, historical, and service quality aspects
2	When two routes have shared path segments, favor the selection of transfer-connections at their merging and splitting transfer-nodes
3	A transfer-connection is relevant if it links stops within a specified threshold of walking-distance (in meters)
4	A transfer-connection is relevant if it complies with a specified threshold of demand (frequency of passengers)

Assumption 2 considers situations where two or more routes share a portion of the path. In these cases, the algorithm prioritizes transfer-connections positioned at strategic stops such as route meeting, merging, or crossing. This assumption translates into a binary variable called *Network strategic value*.

Assumption 3 ensures the geographic vicinity of transfer-connections, mainly to ensure it is walkable. Although a similar criterion was applied in step 2, any connection proposed by experts must also comply with this condition.

Assumption 4 considers that the importance of transfer-connections relates to the number of passengers using them. The implementation of this assumption

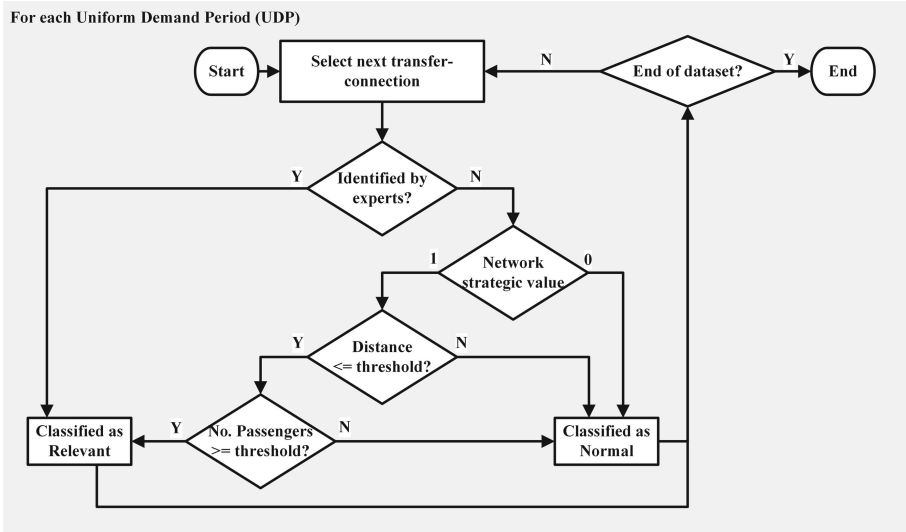


Fig. 7. Algorithm to identify relevant transfer-connections.

consists of selecting all transfer-connections with a frequency of transfer-events higher than a specified threshold.

Figure 7 schematizes the application of the assumptions described in this section into an algorithm. This algorithm runs after the completion of the previous two steps of the methodology.

In contrast to the previous two steps, this algorithm does not use the AFC data-set. The data for this algorithm includes the following features for each transfer-connection in the system: (i) experts identified the connection as relevant (y/n), (ii) the network strategic value (binary), (iii) the walkable distance between the two stops, (iv) the passenger daily frequency. The algorithm performs the sequential validation of all assumptions, as shown in Fig. 7.

4 Results

The database software PostgreSQL was used to select and sort data. The TCM algorithm was implemented in C++, using a 3.4 GHz Intel Core i7 processor and 16 GB of Random Access Memory (RAM). The computational effort of solving the TCM in this particular application is considerably low, less than 10 seconds. The performance of this algorithm in more significant instances was reported in [12]. The methodology described in Sect. 3 was applied to the case study of Porto considering a sample of 4000 randomly selected smart-cards. All smart-cards were analyzed over the entire year of 2013.

This implementation considered the following thresholds: 3 km in assumption 1 of Table 2, 200 m in assumption 2 of Table 3 and in assumption 3 of Table 4, 30 min in assumption 3 of Table 3 and 80 annual transfer-events in assumption 4 of Table 4.

Table 5. Overview of results.

Month	1	2	3	4	5	6	7	8	9	10	11	12	Total
<i>Original data-set</i>													
AFC-records	16783	15080	15965	17160	18460	15088	16349	12891	15232	17851	16360	14719	191938
<i>Alighting not estimated - single daily AFC-record</i>													
AFC-records	1458	1243	1452	1481	1607	1434	1465	1091	1327	1449	1363	1311	16681
%	8.69	8.24	9.09	8.63	8.71	9.50	8.96	8.46	8.71	8.12	8.33	8.91	8.69
<i>Alighting not estimated - distance ≥ 3 km</i>													
AFC-records	447	417	409	475	494	428	404	357	381	464	450	458	5184
%	2.66	2.77	2.56	2.77	2.68	2.84	2.47	2.77	2.50	2.60	2.75	3.11	2.70
<i>Pairs of consecutive AFC-records identified as transfer-connections</i>													
Pairs	1830	1632	1664	1866	2076	1668	1864	1479	1677	1912	1587	1337	20592
%	10.90	10.82	10.42	10.87	11.25	11.06	11.40	11.47	11.01	10.71	9.70	9.08	10.73

Table 5 shows the summary of the results obtained for the first two steps of the methodology as described in Sects. 3.1 and 3.2. The results are detailed by

month, and the last column provides the aggregate value for the entire year. The first row provides information on the total number of AFC records analyzed. The number and percentage of AFC records to which the TCM could not estimate alighting stops are detailed in two groups. The first group includes the cases where there was only one daily record - and therefore, the assumption of returning home could not be applied. The second group refers to the cases where the distance of the estimated alighting stop was higher than 3 km - those estimations were discarded since the passenger is assumed to travel out of the system.

Finally, from the AFC records with successful estimations of alighting stops, Table 5 shows the number of transfer-events that were identified, and its proportion regarding the original AFC data-set. For a yearly aggregate perspective, transfer-events accounted for around 11% of total AFC records. Note that a transfer-event is identified amid two AFC records, but its accounting is not duplicated. Therefore, each transfer-event is accounted for just once - making them comparable to the total number of AFC records.

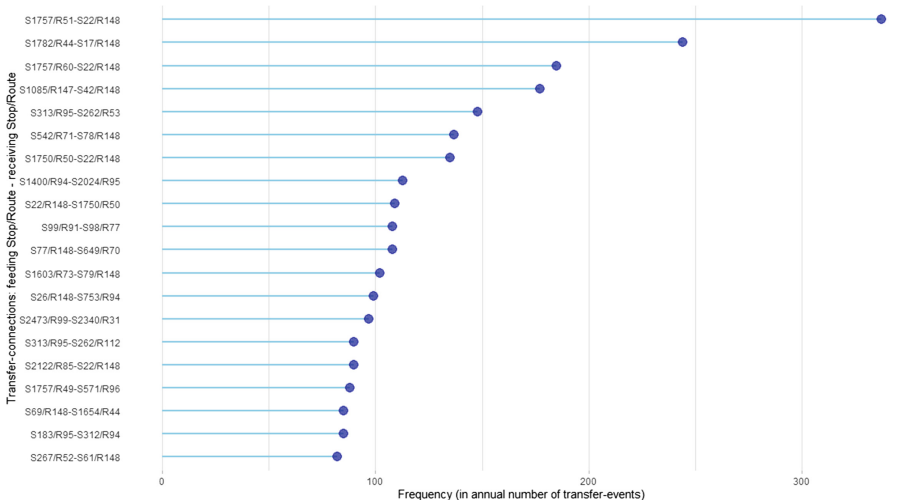


Fig. 8. Selected relevant transfer-connections.

The last step of the methodology proposed in this work was implemented as described in Sect. 3.3. It returned the identification of 20 relevant transfer-connections. Figure 8 shows the selected connections regarding their annual passenger frequency. The selected transfer-connections accounted for 36.40% of all transfer-events under analysis.

5 Conclusions

The main goal of this work was to develop and implement a methodology able to retrieve the most relevant transfer-connections of a PT system. The need to identify the relevant transfer-connections within a PT context arises as a preliminary stage to the implementation of the STP algorithm, usually applied to a simplified network of the PT system which reflects its main demand flows. The identification of such a simplified network is not an easy task, especially in PT systems using entry-only AFC ticketing systems, such as the case study of this work, the city of Porto.

Following this goal, a methodology was developed encompassing three main steps. Step 1 addressed the estimation of the alighting stop of each AFC record using the TCM, step 2 the identification of transfer-events considering all pairs of consecutive AFC records, and step 3 the identification of the most relevant transfer-connections in the PT system following a set of criteria.

This methodology was applied to the case study of Porto, considering a sample of 4000 randomly selected smart-cards over the entire year of 2013. This analysis served as a proof of concept of the methodology. The results obtained are promising and call for the replication of this methodology to larger datasets, and to perform statistic analysis regarding the type of passengers (frequent passenger and occasional passengers), as well as to compare the set of relevant transfer-connections in different UDP, such as peak and off-peak hours of the day. Future work also includes using this methodology to build PT networks of relevant transfer-connections and feed them as inputs to STP algorithms.

Acknowledgements. Funding: This work was supported by (a) the Foundation for Science and Technology (FCT), [grant number PD/BD/113761/2015]; and (b) by the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the FCT within project POCI-010145-FEDER-032053 - PTDC/ECI-TRA/32053/2017.

References

1. Ceder, A.: *Public Transit Planning and Operation: Theory, Modeling and Practice*. Elsevier, Butterworth-Heinemann (2007). ISBN 978-0-7506-6166-9
2. Ibarra-Rojas, O.J., Rios-Solis, Y.A.: Synchronization of bus timetabling. *Transp. Res. Part B: Methodol.* **46**(5), 599–614 (2012). <https://doi.org/10.1016/j.trb.2012.01.006>
3. Fouilhoux, P., et al.: Valid inequalities for the synchronization bus timetabling problem. *Eur. J. Oper. Res.* **251**(2), 442–450 (2016). <https://doi.org/10.1016/j.ejor.2015.12.006>
4. Cao, Z., et al.: Optimal synchronization and coordination of actual passengerrail timetables. *J. Intell. Transp. Syst.* **23**(3), 231–249 (2019). <https://doi.org/10.1080/15472450.2018.1488132>
5. Barry, J., et al.: Origin and destination estimation in New York City with automated fare system data. *Transp. Res. Rec.: J. Transp. Res. Board* **1817** (2002). <https://doi.org/10.3141/1817-24>

6. Trépanier, M., Tranchant, N., Chapleau, R.: Individual trip destination estimation in a transit smart card automated fare collection system. *J. Intell. Transp. Syst.: Technol. Plan. Oper.* **11**(1), 1–14 (2007). <https://doi.org/10.1080/15472450601122256>
7. Barry, J., Freimer, R., Slavin, H.: Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. *Transp. Res. Rec.: J. Transp. Res. Board* **2112**, 53–61 (2009). <https://doi.org/10.3141/2112-07>
8. Alsgar, A., et al.: Use of smart card fare data to estimate public transport origin-destination matrix. *Transp. Res. Rec.: J. Transp. Res. Board* **2535**, 88–96 (2015). <https://doi.org/10.3141/2535-10>
9. Nunes, A.A., Galvão, T., Cunha, J.F.: Passenger journey destination estimation from automated fare collection system data using spatial validation. *IEEE Trans. Intell. Transp. Syst.* **17**(1), 133–142 (2016). <https://doi.org/10.1109/TITS.2015.2464335>
10. Munizaga, M., Palma, C.: Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transp. Res. Part C: Emerg. Techn.* **24**, 9–18 (2012). <https://doi.org/10.1016/j.trc.2012.01.007>
11. Nassir, N., et al.: Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. *Transp. Res. Rec.: J. Transp. Res. Board* **2263**(1), 140–150 (2011). <https://doi.org/10.3141/2263-16>
12. Hora, J., et al.: Estimation of Origin-Destination matrices under Automatic Fare Collection: the case study of Porto transportation system. *Transp. Res. Procedia* **27**, 664–671 (2017). <https://doi.org/10.1016/j.trpro.2017.12.103>