Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation

António A. Nunes, Teresa Galvão Dias, and João Falcão e Cunha, Senior Member, IEEE

Abstract—A methodology for estimating the destination of passenger journeys from automated fare collection (AFC) system data is described. It proposes new spatial validation features to increase the accuracy of destination inference results and to verify key assumptions present in previous origin–destination estimation literature. The methodology applies to entry-only system configurations combined with distance-based fare structures, and it aims to enhance raw AFC system data with the destination of individual journeys. This paper describes an algorithm developed to implement the methodology and the results from its application to bus service data from Porto. The data relate to an AFC system integrated with an automatic vehicle location system that records a transaction for each passenger boarding a bus, containing attributes regarding the route, the vehicle, and the travel card used, along with the time and the location where the journey began. Some of these are recorded for the purpose of allowing onboard ticket inspection but additionally enable innovative spatial validation features introduced by the methodology. The results led to the conclusion that the methodology is effective for estimating journey destinations at the disaggregate level and identifies false positives reliably.

Index Terms—Automated fare collection, O-D matrix, public transport, spatial validation, travel patterns.

I. INTRODUCTION

A

UTOMATED fare collection (AFC) systems are used in many urban public transport systems around the world. As the designation suggests, these are typically designed with the specific purpose of automating the ticketing system, easing public transport use for passengers and adding efficiency to revenue collection operations. In addition, AFC systems are used to enable integrated ticketing across different public transport modes and operators in urban areas. While these are their primary design functions, AFC systems continuously generate data that may be useful for service performance monitoring and for decision-making support [1], [2]. The work described in this paper aims to use raw AFC system data to estimate the destination of individual passenger journeys.

Two main configurations of AFC systems exist, depending on whether passenger fare media are read just at the beginning or both at the beginning and end of journeys. The first of these are known as entry-only AFC systems and require additional logic for estimating the destination of passenger journeys because alighting locations are not recorded [3]. Entry-only configurations are popular in bus services around the world, designed to avoid alighting delays if the fare media of all exiting passengers had to be read upon arrival at a stop. Since alighting locations are not recorded, flat fare structures have often been used to lessen the need for on-board inspection to control underpaid travel. However, exceptions to this are becoming increasingly common, as public transport agencies are driven to deliver more equitable distance-based pricing. Hence, the motivation of this work is the development of a methodology to estimate the destination of passenger journeys using entry-only AFC system data with a distance-based fare structure. The specific characteristics of these systems enable a spatial validation feature introduced by the methodology. It consists of a comparison, at an individual journey basis, between the estimated travel distance and the paid fare.

The goal of the methodology is to enrich raw AFC system data into complete Origin-Destination (O-D) passenger journey data sets depicting individual travel patterns. This requires high precision from the estimates and results at maximum disaggregation level, so the methodology favors accuracy over the percentage of inferred journey destinations. The resulting information is useful for adjusting public transport offerings to passenger demand and allows the construction of O-D matrices at any level of aggregation and geographic coverage. The methodology was applied to the Andante system in Porto as a case study, using data from its main bus service operator called Sociedade de Transportes Coletivos do Porto, SA (STCP), which runs the vast majority of routes within the city and into the surrounding metropolitan areas. A specificity of Andante is that it is a time-based system for users without a fixed subscription, which favors another validation feature introduced by the methodology. It deals with a comparison between the estimated geographic coverage of a journey and the location of duplicate transaction records made by passengers checking remaining travel time. The introduction of the aforementioned spatial validation features, relating to a comparison between travel distance and paid fare, and to the location of duplicate transaction records, is one of the main contributions of this work. The other is the identification of single daily journeys with multiple stages
for reducing inference errors. The results obtained suggest that the methodology is effective for estimating the destinations of journeys at disaggregate level and reliable in the detection of false positives. The new spatial validation features suggest that the key assumptions present in previous literature in the field are largely valid for the Andante case.

This paper is structured as follows. Section II is a review of related works, including their case studies, objectives, and validation techniques used. Section III presents the case study, available data, and its preparation. Section IV describes the methodology, and Section V details its implementation. Section VI discusses the results. Lastly, Section VII provides conclusions and outlines future work.

II. RELATED WORK

The estimation of O-D matrices from entry-only AFC system data is a topic that has received substantial research interest. Reliable O-D information is vital for the monitoring and planning of public transport systems as it depicts travel demand, but is generally challenging to obtain. Destination estimation and O-D matrix estimation are not the same issue, but the designations are often used interchangeably. The difference falls upon the level of data aggregation. The objective of O-D matrix estimation is to determine aggregate travel counts between O-D pairs, whereas destination estimation looks at each journey individually. However, most O-D matrix estimation works rely on a destination estimation algorithm, which outputs are subsequently aggregated. Hence, the subjects are intertwined.

The creation of O-D matrices traditionally relies on extensive travel surveys carried out in a periodic basis, which are expensive to conduct and prone to response bias [4]. Barry et al. proposed a methodology to overcome both of those problems with a set of algorithms applied to entry-only AFC data from the New York City subway system [4]. Their approach introduced two seminal assumptions that have been applied in several studies afterward, and which are based on the fact that passenger journey origins are known from entry-only AFC data, but destinations are not. The first of these assumptions is that the most likely destination of a passenger journey is the origin of the next journey. The second is that the most likely final daily destination of a passenger is the first daily origin. From these assumptions they were able to infer the destination for 83% of boarding transactions in a single day sample. Barry et al. later expanded this approach including both New York City subway and bus data [3]. The addition of bus data required a slight modification to the aforementioned assumptions, to consider that an estimated destination may not be the same but the nearest stop to the related origin.

Evidently these assumptions do not always hold. A public transport passenger may travel an intermediate journey segment on foot or by car for instance, which will break the assumption of the destination being nearest to the next origin. Similarly, if a public transport passenger stays in different places overnight the second assumption of the final daily destination being nearest to the first origin will likely be broken. Therefore, the validity of results obtained from O-D estimation methodologies should be verified. Barry et al. validated their methodology using travel diary information and found that the assumptions held for 90% of users surveyed [4]. These authors compared their inferred destination totals with exit counts at the stations, and estimated peak load point passenger volumes to increase the robustness of their assessment. This sort of validations has been called exogenous because they rely on external data sets instead of the actual AFC system data to which the assumptions were applied [5].

The aforementioned assumptions [4] have been used in later research work on the topic. Trépanier et al. applied a comparable methodology to bus service data from Gatineau, Québec, but introduced an endogenous validation step with their methodology [6]. It required the candidate destination to be within a 2000 m Euclidean distance from the related origin, or else it was assumed that an intermediate journey segment in an unrecorded mode of transport might have taken place. Another difference is how boarding transactions that are single in a day for a particular passenger were dealt with; instead of not inferring the destination for those records, the authors opted to carry out a journey regularity analysis for the passenger and estimate the destination on that basis whenever enough data was available [6]. They were able to infer the destination for 66% of journeys. Zhao et al. use a comparable methodology on Chicago Transit Authority rail system data, but with a stricter candidate destination cutoff Euclidean distance of only 400 m [2]. Although these authors did not attempt to infer destinations of single day journeys, they were still able to infer the destination of 71% of journeys over a six-day period, and their method was partially validated at aggregate level using O-D survey data.

Further studies have applied similar logic. Farzim applied an O-D estimation methodology to bus system data in São Paulo and attempted to validate the results with an O-D household survey [7]. However, the conclusions were hampered by the scarcity of buses equipped with Automated Vehicle Location (AVL) technology and by a time gap between data sets. A methodology proposed by Li et al. using bus data from Jinan, China, is claimed to have inferred the destination of 75% of journeys, but appears to rely exclusively on a transfer distance based endogenous validation [8]. Wang et al. have applied their methodology to AFC bus system data from Transport for London (TfL) combining two types of validations [1]. The first consisted of a maximum transfer Euclidean distance, whereas the second was a comparison of the results obtained from applying their methodology with an extensive bus O-D survey, which yielded promising results. Their method did not attempt to infer destinations of single day journeys, and was able to infer the destination of 57% of journeys. Drawing from several of the previous methodologies, Gordon proposed a sophisticated algorithm to infer both the destination and time of arrival of bus passenger journeys that was similarly tested with TfL data [9]. Inclusion of the time aspect enabled an additional validation if the passenger had enough time to transfer on foot or not. Destinations were inferred for 74% of journeys.

Munizaga and Palma propose a methodology with a slight variation that considers generalized time (a combination of walking time and vehicle travel time) in addition to distance in determining the potential destination of a journey [10]. It has been applied to AFC bus and metro system data in Santiago
and obtained around 80% inference rate in two weekly samples considering a maximum transfer Euclidean distance of 1000 m. The methodology was later used by Devillaine to estimate the location, time, and purpose of activities of public transport users [11]. Munizaga et al. later built on that methodology by proposing more robust endogenous validations methods, and presented encouraging results from exogenous validation at a disaggregate level from an O-D survey and an experiment with volunteers [5]. Recent research has sought to expand the use of AFC system data beyond destination inference to deduct route choices [12]. Another new stream of research is automated passenger tracking, which has shown to be a promising addition to entry-only systems for recording alighting location without egress gates [13]. Lastly, it is noted that the estimation of O-D matrices is a topic of interest in other transport domains, including private vehicular transport [14], [15], freight transport [16], and pedestrian movement [17], using count data sources or mobile phone traces [18], [19].

In summary, previous work on the topic has focused on a number of urban public transport systems around the world, all of which have their own specificities regarding travel behavior, data set availability, and degree of integration between systems. The proposed methodologies equally have specificities, namely in terms of the variety and strictness of validation rules that are aligned with their particular goals. Hence, the outcomes are varied and not directly comparable, but there is general belief that the main assumptions are valid in the majority of instances. However, all of these works have been applied to entry-only systems with flat fare structures, hence no previous attempts have been made to use additional data resulting from the operation of distance-based fares to increase the accuracy of the destination inference results. Increasing the accuracy raises confidence that destinations are correctly inferred, yet decreases the overall inference rate.

III. ANDANTE SYSTEM

Andante is an entry-only AFC system with a distance-based fare structure that covers the metropolitan area of Porto. Similar systems exist to which this work applies, such as the Leap Card in Dublin buses and the SL Access in Stockholm buses. In addition, Toronto Transit Commission and Utah Transit Authority are considering switching to distance-based fares, whereas First buses in the West of England have already made it. Distance-based fare structures are likely to become increasingly common because they deliver fairer pricing for users. Although the proposed methodology applies to AFC systems of similar characteristics, each will have specificities that must be understood and considered for the extraction and preparation of data.

The fare media of Andante are contactless travel cards that can be used on buses from several operators, as well as on metro and railways. Distance-based fares are defined by a zonal structure. The system is divided into geographic travel zones and the journey fare depends on the number of zones traveled between its origin and destination. Andante is a time-based system, allowing pay-per-use passengers to make unlimited transfers in a given time period, which increases according to the number of zones that are included in the respective fare. A journey relates to a single fare and consists of one or more journey stages in different routes or vehicles. The Andante system creates a transaction record every time a passenger taps a travel card on a reader. This must happen at the beginning of each journey stage, when changing routes or entering another vehicle. However, being a time-based system, pay-per-use passengers—those without a monthly subscription—sometimes tap their cards during a journey to check on the display of the reader how much time is left for traveling without paying another fare. In addition, this may happen when a passenger cannot remember tapping the card and repeats it in the same journey stage.

Each Andante transaction record contains several data attributes, of which the following are of interest to the proposed methodology:

1. Travel card serial number;
2. Station or bus stop where the transaction took place;
3. Route number (only applicable to bus journeys);
4. Direction of travel (only applicable to bus journeys);
5. Vehicle number (only applicable to bus journeys);
6. Vehicle trip start time (only applicable to bus journeys);
7. Transaction timestamp; and
8. Number of travel zones in the travel card.

In the case of buses, although card readers are located inside the vehicle, the stop where the transaction took place is recorded because the Andante AFC system is fully integrated with an AVL system. The integration of the AFC and AVL systems spares the need to infer journey origins as in many cases found in the O-D estimation literature (e.g., [1], [2], [7]). The AVL system is believed to assign the transaction stop code with high accuracy; it is used to inform the passengers of the next bus stop and has been observed to perform that task very reliably and consistently. Some of the aforementioned data attributes are recorded for the main design purpose of allowing on-board ticket inspection, but are useful for estimating the destination of journeys.

A. Available Data

The data used to illustrate the methodology is the set of transaction records in the whole month of April 2010 in STCP buses. April 2010 was a normal month with 21 working days and 20 school days. All STCP buses were fitted with AVL equipment, which is a great benefit for testing the methodology. At the time there were two types of travel cards, the Andante multimodal and the STCP specific. The latter has since been disused and contained fewer usable data attributes, therefore only the Andante travel card user transactions were considered, totaling approximately three million records for that period. This subset of data contains fewer travel card serial numbers, but maintains the entire time period for each one of them, which is well suited for illustrating travel patterns. Fig. 1 depicts the geographic distribution of boarding transactions in the data subset and the Andante zonal structure. The methodology requires the Andante AFC system data to be fused with three additional data sources, which in the case of Andante are publicly available. The first is the listing of bus stops (2377 in total), along with their stop code, zone, and location coordinates. The second
is the structure of bus routes (67 in total), listing the bus stops and their respective sequence in each direction of travel. The third is a matrix with the minimum number of zones traveled between O-D zone pairs (for a total of 18 zones).

B. Data Preparation

Some authors have previously noted that AFC system data often contains transaction records with missing or illogical attribute values [3], [6]. Hence, it is necessary to prepare the data beforehand. The Andante system data set is not an exception. The proportion of transaction records in the data sample missing at least one of the relevant data attributes was initially 3.2%. Additionally, 0.8% of transaction records were found to have illogical values across two attributes. However, it has been possible to identify and, in 81.5% of those cases, correct such errors from thorough analysis and subsequent pre-processing of data.

The most frequent cause of missing and illogical attribute values of the Andante data is the changeover between consecutive vehicle trips. Upon arrival at the terminus stop, the bus driver needs to signal completion of the trip and afterward signal the beginning of the next trip if not returning the vehicle to the depot. That next trip often is a return in the same route, but in the opposite direction. The data sample shows that passengers sometimes board the bus, initiating their journey before the changeover process is completed. Two different types of error occur in those circumstances. The first is when a passenger boards the bus before the driver signals the completion of the previous trip. This creates an illogical value in terms of the bus stop code attribute of that transaction record, because it fails to make sense that a passenger would board the bus at the terminus stop. The second is when a passenger boards the bus after the driver signals the completion of the previous trip, yet before the next trip is initiated. This originates missing values for the direction of travel and vehicle trip start time attributes of that boarding transaction record. Both types of errors originating from the changeover process were mitigated by assuming with a high level of confidence that those boarding transactions should instead be assigned to the next trip in that vehicle. Both the direction of travel and vehicle trip start time can be assigned to those records based on boarding transaction records created later on that bus trip, which can be uniquely identified from the data attributes vehicle number and vehicle trip start time.

The data sample contained few cases of transaction records missing the direction of travel attribute value. It is unclear why that information is missing, but such errors were easily mitigated by checking the direction of travel from other transaction records in the same trip. The application of the mitigation measures that were mentioned reduced the proportion of transaction records in the data sample with missing or illogical values to 0.7%. This breaks down as shown in Table I into transaction records missing the bus stop code or the vehicle trip start time, and transaction records at unknown bus stop codes or at bus stop codes that are not part of the bus route. The missing bus stop codes are likely the result of a communication failure of the AVL system, whereas the causes for the remaining cases are not entirely clear, but are not a significant share of the data sample. The transaction records that remain with missing or illogical data attributes after pre-processing are not discarded from the data sample otherwise travel patterns would become distorted due to lost journey stages. This approach will improve the overall accuracy of estimates.

IV. DESTINATION INFEERENCE METHODOLOGY

The objective of the methodology for estimating the destination of passenger journeys from entry-only AFC system data is to determine the alighting stop of each journey stage:

\[ \hat{s}_{pjk}^a, \forall p, j, k \]

where,

\[ \hat{s}_{pjk}^a \quad \text{estimated alighting stop of the } j\text{th journey stage of passenger } p \text{ on day } k. \]

The word stop is used throughout this section to generalise rail stations and bus stops. Indices \( p, j, \) and \( k \) are used with the same meaning throughout the upcoming equations. The methodology is primarily based on the two key assumptions found in the literature reviewed earlier [3], [4] relating to the continuity of daily travel and to the circularity of daily journey chains. Additionally, the direction of travel is known for all transaction records in the Andante bus data set, hence it can be said that the destination of a journey stage must be downstream
from its origin in this case. Let a public transport route in a given direction and the set of candidate alighting route stops of a passenger journey stage be respectively defined as:

\[ R = \{ s_i^R | 1 \leq i \leq n^R \} \]  \hspace{1cm} (2)

\[ A_{pjk}^R = \{ s_i^R | \beta_{pjk} \leq i \leq n^R \} \]  \hspace{1cm} (3)

where,

- \( i \) stop sequence index;
- \( R \) route \( R \) in a given direction;
- \( s_i^R \) \( i \)th stop of route \( R \);
- \( n^R \) number of stops of route \( R \);
- \( A_{pjk}^R \) set of candidate alighting stops along route \( R \) of the \( j \)th journey stage of passenger \( p \) on day \( k \);
- \( \beta_{pjk} \) boarding stop sequence of the \( j \)th journey stage of passenger \( p \) on day \( k \).

It is noted that in rail-based systems the direction of travel is usually unknown because card readers are located outside vehicles in stations that cater for various directions. The reduction of the number of candidate alighting stops cannot be performed in those cases. Furthermore, the wording would be slightly different if the methodology were to be applied to rail-based modes or a multimodal public transport system, but the basis would remain the same. In this case, the key assumptions can be summarized in the following way:

1) The most likely destination of a journey stage is the route stop located downstream from its own origin that is nearest to the origin of the next journey stage from that passenger.

2) The most likely destination of the last journey stage of a day is the route stop located downstream from its own origin that is nearest to the origin of the first journey of the day from that passenger.

Let \( d(x, y) \) be the Euclidean distance between route stops \( x \) and \( y \), the key assumptions are respectively formulated as:

\[ \hat{s}_{pjk}^a \leftarrow \text{arg min}_{s_{pjk}^a} \left\{ d \left( s_{p(k+1)k}^b, s_{pjk}^a \right) , s_{pjk}^a \in A_{pjk}^R \right\} , j < m_{pk} \] \hspace{1cm} (4)

\[ \hat{s}_{pjk}^a \leftarrow \text{arg min}_{s_{pjk}^a} \left\{ d \left( s_{p1k}^a, s_{pjk}^a \right) , s_{pjk}^a \in A_{pjk}^R \right\} , j = m_{pk} \] \hspace{1cm} (5)

where,

- \( s_{pjk}^a \) alighting route stop candidate of the \( j \)th journey stage of passenger \( p \) on day \( k \);
- \( s_{pjk}^b \) boarding route stop of the \( j \)th journey stage of passenger \( p \) on day \( k \);
- \( m_{pk} \) number of daily journey stages of passenger \( p \) on day \( k \).

Applying the key assumptions sets out a candidate destination for each boarding transaction record (Fig. 2), unless there is a single daily journey for a passenger \( (m_{pk} = 1) \) in which case the candidate of its last stage is not determined. It was decided that the destination of single daily journeys would not be inferred because if a passenger occasionally changes part of the daily routine, assigning a destination based on past behavior will return unreliable results. Given the objective of this work to illustrate travel patterns, requiring focus on the individual rather than aggregate O-D matrix estimates, the adopted approach avoids adding bias to the travel history of passengers.

After setting candidate destinations, spatial validation rules are used to ascertain whether these assumptions are likely to hold for each individual transaction record. The methodology proposes four endogenous spatial validation rules that can be described by the following questions. The third and the fourth are newly introduced (the fourth relates exclusively to bus journeys):

1) Are the origin and candidate destination of a journey stage the same?

2) Is the candidate destination of a journey stage beyond a set Euclidean distance from the next journey origin (or from daily origin if stage is last) for that passenger?

3) Is the number of travel zones exceeded for the passenger to reach the candidate destination?

4) When a travel card is tapped on the reader more than once from daily origin if stage is last) for that passenger?

The first rule has the purpose of verifying, as applicable, if the origins of two consecutive boarding transaction records from a passenger are at the approximate same location, or if the origins of the first and last boarding transaction records of the day from a passenger are at the approximate same location [10]. This is equivalent to checking the equality:

\[ s_{pjk}^b = \hat{s}_{pjk}^a \] \hspace{1cm} (6)

If the answer is yes, it is hypothesized that a return journey segment happened in another mode of transport, because it makes no sense to board a vehicle and not go anywhere. Consequently, the destination of that journey stage is not inferred.

The second rule evaluates the likelihood of the candidate destination being the actual destination of a journey stage based on walking distance [6]. This is equivalent to checking one of the following inequalities, depending on the passenger journey being the last of the day or not:

\[ d \left( \hat{s}_{pjk}^a, s_{p(j+1)k}^b \right) > c, j < m_{pk} \] \hspace{1cm} (7)

\[ d \left( \hat{s}_{pjk}^a, s_{p1k}^a \right) > c, j = m_{pk} \] \hspace{1cm} (8)
where, $c$ is cutoff Euclidean distance. The cutoff Euclidean distance can be parameterized and has been set in this case at 640 m, which is intentionally shorter than found in most of the previous literature. This deliberately makes the present approach relatively conservative in terms of its endogenous distance-based validation to suit the main objectives of the work, but should be appraised in light of the density of the urban fabric and of the public transport system under analysis. Moreover, this parameter value is not arbitrary. It comes from the Public Transport Accessibility Levels (PTAL) methodology, representing a walk catchment area of 8 min at 4.8 km/h assumed to be the longest distance a person would normally walk to access a bus service [20]. Sensitivity of the cutoff distance parameter is discussed in Section VI. Therefore, if the answer to the second question is yes, it is assumed that an intermediate journey segment took place in another mode of transport and the destination of that journey stage is not inferred.

The third spatial validation rule relates specifically to entry-only AFC systems with a distance-based fare structure such as Andante, which is divided into travel zones. Other systems exist that use stages instead of zones, but a similar principle applies. Since the number of travel zones in a card is known in this case, and assuming that passengers do not travel beyond the zones they are legally entitled to, it is possible to validate if the candidate destination falls within the allowable travel bounds. This is equivalent to checking the inequality:

$$z_{pjk} \leq z^{\min}(s^a_{pjk}, s^b_{pjk})$$

where,

$$z_{pjk}$$ number of zones in travel card of passenger $p$ in the $j$th journey stage on day $k$;

$$z^{\min}(s_x, s_y)$$ minimum number of zones between stop $x$ and stop $y$.

The assumption is argued to be reasonable in the context of Andante because fraud levels of any kind are very low totaling 0.49% of inspected passenger journeys [21], which is likely a result of frequent on-board inspections and of the penalty applied to underpaid travel being 100 times the fare. A positive answer to the third question indicates that the key assumptions do not hold for that journey and its destination cannot be inferred.

Lastly, the fourth spatial validation rule owes to Andante being a time-based system for pay-per-use passengers and therefore it may or may not apply to other AFC systems. Such passengers occasionally tap their cards on a reader during a journey because it provides feedback on how much time is left for traveling without paying another fare. There may be other cases where passengers do that having forgotten whether they tapped the card entering the vehicle or not. Either way, it creates a duplicate transaction record for that journey stage. This only applies to bus journeys, since in rail-based modes an additional transaction record is inevitably interpreted as a new journey stage. This methodology identifies such duplicate records and uses them to validate if the duplicate transaction happens downstream from the candidate destination in that bus route. This is equivalent to checking the inequality:

$$\rho_{pjk} > \delta_{pjk}$$

where,

$\rho_{pjk}$ highest route stop sequence of duplicate records in the $j$th journey stage of passenger $p$ on day $k$;

$\delta_{pjk}$ estimated alighting route stop sequence of the $j$th journey stage of passenger $p$ on day $k$.

If the inequality is verified, the answer to the fourth question is positive and the destination cannot be inferred because the passenger was confirmed to be traveling in the same bus trip beyond that location.

V. IMPLEMENTATION

Fig. 3 illustrates the algorithm developed in SQL to implement the proposed methodology. It has linear complexity; the execution time is proportional to the number of transaction records selected. The algorithm goes through the transaction record data set sorted firstly by travel card serial number and secondly by the travel card transaction timestamp. The first decision is to verify if the record is a duplicate, in which case it will be used for spatial validation, but its destination shall not be inferred.

The following decision is to check whether the record is the last or the only stage of a single daily journey on that travel card serial number. Two aspects are highlighted here. The first is the day interval definition. In this case, the data set reveals significant transaction levels around midnight, dropping steadily to minimums between 3:00 am and 5:00 am. Transactions between midnight and 3:00 am appear to be largely related to the previous day passenger journey chains. Coincidently, the shift from night time to daytime bus services happens around 5:00 am. Therefore, it was decided that in terms of daily journey chains a day starts at 5:00 am and ends at 4:59 am of the following morning.

The second aspect relates to the distinction between journey stages and complete journeys. Passengers often have to change between public transport routes to reach their destination and, in the case of the Andante system, tap their travel card on a reader every time they board a different vehicle. Each of those transactions relate to a stage of their complete journey. The difference matters in cases when there is a single daily journey for a passenger. If that journey is single staged, it is trivial that a destination cannot be inferred due to lack of information to determine a candidate destination. But if that journey has several stages, it is arguably most likely that the last journey stage was to reach a destination other than the daily origin (left in Fig. 4), otherwise the passenger would be traveling in a circle (right in Fig. 4). Both of these scenarios are possible in theory, but simply assuming that the latter is true carries great risk of inferring the destination incorrectly. Therefore, a candidate for that destination is not determined either to meet the objective of highest accuracy of estimates, instead of assuming it to be the daily origin as seen in previous literature. It is reminded that the stages of a complete journey are defined by the time-based Andante rules for pay-per-use passengers, which set out maximum journey durations according to the number of travel zones.
The following decision (Fig. 3) is to check if the transaction record is the last stage of the day for that travel card serial number. This determines which of the key assumptions should apply for setting the candidate destination. The next step is to verify if the origin and candidate destination route stop codes are both present in the transaction record and are logical, or else its destination cannot be inferred. This cannot be fulfilled earlier, since the candidate destination is only determined after the application of the key assumptions. If the necessary data attributes are present, the record goes through the four endogenous spatial validation rules already described, which deal respectively with the inequality of origin and destination, the maximum interchange distance, the adequacy of the number of zones in the travel card, and the inexistence of duplicate records downstream from the candidate destination. When a record survives all four spatial validation rules, its destination is subsequently inferred with great confidence.

Although the inference of arrival time is not paramount for the purposes of this work, the next step, applicable to bus services, is to verify if there are boarding transactions in the same service at the inferred destination stop. If there are, the first timestamp of such transactions is assigned as the arrival time. If not, an additional step verifies if there were boarding transactions in the same bus service simultaneously between the origin and the inferred destination, and after the inferred destination. The presence of such transactions allows the definition of arrival time upper and lower bounds, and the interpolation of an estimated arrival time using the number of stops in between as a weighing factor. The absence of enough boarding transactions in the same bus service renders the estimation of the arrival time unfeasible without exogenous AVL data.

VI. DISCUSSION OF RESULTS

The algorithm was executed in a computer with a 1.6 GHz Intel Core i5 processor and 4 GB of Random Access Memory (RAM). The results were obtained for the whole month of April 2010, but with live data it can be run incrementally on a daily basis in less than 10 min for the magnitude of this case. Application of the methodology to the Andante AFC system bus data yielded a percentage of inferred destinations of 62.4% of all transaction records. This result is partially dictated by the nature of the data from Porto STCP buses, particularly in terms of the amount of single daily journeys that are an expression of travel behavior. But it is heavily influenced by
the application of strict validation rules to identify individual records for which the key assumptions—continuity of daily travel and circularity of daily journey chains—do not hold. To put these results into perspective, if the endogenous spatial validation rules described in Section IV were not applied, and if the proposed distinction between journey stages and complete journeys were ignored, the methodology would yield a percentage of inferred destinations of 81.6% using the same data set. However, the accuracy of the additional estimates (almost 20% of the total) would be compromised and the methodology would not reflect the objectives of this work. This highlights one of its contributions, of avoiding inference errors by identifying and dealing cautiously with single daily journeys with multiple stages. This approach favors the highest accuracy of estimates at disaggregate level over the percentage of inferred journey destinations.

Table II breaks the results down. The boarding transactions that survive validation are 62.4% of the total, and subdivide into those where the arrival time was inferred or bound (36.6% of total), and those for which there was not enough information to estimate an arrival time (25.8% of total). The remaining transaction records total 37.6%. The majority are either the only or the last stage of single daily journeys (21.2% of total), many fail spatial validation (15.1% of total), and some have data attribute errors (0.9% of total) or are duplicates (0.4% of total). The implementation of the methodology applies the spatial validation rules in the specified order, otherwise some transaction records could fail more than one validation simultaneously. An example would be a record for which the maximum interchange distance was exceeded and so were the number of zones in the travel card. In such case the former spatial validation is considered the cause for the failure in inferring a destination. It is noted that the percentage of records without an inferred destination could be decreased if the data set available were multimodal. It would reduce the number of records that are the only or last stage of single daily journeys, or that fail the first validation rule, as further journeys may exist for those passengers within the Andante system with other operators.

Additional results were obtained from applying the methodology to available data from a previous month. The objective has been to check for significant variations, yet the results were very similar. Using data for the whole month of January 2010, the percentage of boarding transactions that survived validation were 62.1% of the total (−0.3%), which subdivided into 35.8% of the total (−0.8%) having the time inferred or bound, and 26.3% of the total (+0.5%) not having enough information to estimate an arrival time. The remaining transaction records from January 2010 totaled 37.9% (+0.3%).

Sensitivity analysis of the cutoff Euclidean distance parameter is shown in Table III. If it were made stricter by reducing it from 640 m to 400 m as in [2], the percentage of transaction records failing the second spatial validation rule would increase to 10.5% and the percentage of inferred destinations would drop to 60.0%. Conversely, if the parameter were made more liberal by increasing it to 1000 m as in [10], those percentages would respectively drop to 5.8% and increase to 64.6%. The variation is significant, however the parameter set was felt to be an adequate trade-off between the risks of rejecting true positives (rejecting a correct candidate destination) that is greater with a short cutoff distance, and accepting false positives (accepting an incorrect candidate destination) that is greater with a longer cutoff distance.

Another contribution of this work is the introduction of the third and fourth spatial validation rules with the main purpose of testing at the maximum disaggregation level—single journey stage—that the two key assumptions underlying most O-D estimation methodologies in the literature apply to a particular case study. Each public transport system has specific usage patterns; therefore the fit of key assumptions may vary and should be tested for each particular case. Previously those assumptions had mostly been validated at an aggregate level using external data from O-D surveys and exit counts. However, there is a risk concerned with the errors tending to average out when the assumptions are tested at aggregate level and there is no bias. For example, if the assumptions were flawed, an overestimate of journeys from origin A to destination C may compensate an underestimate of journeys from origin B to destination C. Despite this risk, endogenous and exogenous validation rules should ideally have been combined. It was not possible for the present case study due to the inexistence of a recent O-D survey or Automatic Passenger Count (APC) devices in STCP buses. The latter had reportedly been trialed, but the operator found the APC technology unreliable and decided against its use.
As shown in Table II, the newly introduced spatial validation rules supported the validity of the key assumptions for the vast majority of journeys of the Porto STCP buses case study. In fact, the percentage of boarding transaction records that fail the third and fourth spatial validation rules is reduced. Particularly the fourth spatial validation rule that deals with duplicate transactions may at first seem unnecessary for the present case study, since it highlights less than 0.1% of boarding transaction records that survived previous validation rules. But this result must be read with caution. One should look at the percentage of boarding transaction records that fail the duplicate record validation rule in relation to those where the destination has been inferred and have at least one duplicate: 61 out of 6155, approximately 1.0%. This low value provides evidence that no reasons exist to believe the key assumptions are not valid for the vast majority of cases. This claim applies as long as the first and second spatial validation rules dealing with the inequality of origin and destination, and with the maximum interchange distance are carried out beforehand.

Lastly, it is noted that the methodology estimates the destination of passenger journeys at maximum disaggregation level, which allows for the construction of O-D matrices at any desired level of aggregation and geographic coverage. The results show that complete journeys are composed by 1.2 stages on average.

VII. Conclusion and Future Work

This paper described a methodology for estimating the destination of passenger journeys from AFC system data. It builds on previous work found in the literature by replicating key assumptions, but introduces a methodology that is specifically applicable to the case of entry-only systems with a distance-based fare structure, which had not been addressed before. The proposed methodology makes two contributions. First, it proposes new endogenous spatial validation rules at disaggregate level. These additional validation rules deal with the number of zones or stages in a travel card—which is specific to distance-based fares—and with the existence of duplicate transaction records. Their purpose is to test the validity of key assumptions regarding continuity of daily travel and the circularity of daily journey chains, on a single case basis and at maximum disaggregation level. For the Porto STCP buses case study, the spatial validation rules were not prolific in the identification of false positives that were unspotted from previous validation steps, but did support the validity of the key assumptions. The second contribution relates to improved reliability of estimation results. The methodology refines previous work by distinguishing between journey stages and complete journeys and subsequently not inferring the destination of the last stage of single daily passenger journeys with multiple stages. Such instances otherwise introduce a great deal of uncertainty to the estimation results.

This work introduced AFC system data from the main bus operator in Porto as a new case study to the O-D matrix estimation literature. The methodology proved effective to estimate the destination of journeys at disaggregate level and to detect instances where the candidate destination obtained from the application of key assumptions is likely incorrect. The approach toward these instances is conservative; their destinations are not inferred. The percentage of inferred destinations is largely influenced by the nature of data from Porto STCP buses and by the strictness of validation rules seeking the highest accuracy of estimates. Future work will focus on exogenous validation of the methodology once up-to-date O-D survey results become available from STCP. Future improvements to the methodology may include an additional validation rule based on an interchange time interval. Additionally, it is expected that new multimodal AFC system data will become available in the near future, which will allow the application of the methodology to boarding transactions across all Andante operators.

ACKNOWLEDGMENT

The authors thank STCP for providing the AFC system data necessary for this work.

REFERENCES


António A. Nunes received the M.Eng. in industrial engineering and management from the University of Porto, Porto, Portugal, in 2004 and the M.Sc. degree in urban studies from University College London, London, U.K., in 2010. He is currently working toward the Ph.D. degree in transportation systems with the Faculty of Engineering, University of Porto, and a researcher at Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC). His research interests include intelligent transport systems, spatiotemporal data mining, origin-destination matrices, and social network analysis.

Teresa Galvão Dias received the M.Sc. degree in mathematics from the University of Coimbra, Coimbra, Portugal, in 1990 and the M.Sc. degree in electrical engineering and computers and the Ph.D. degree in sciences of engineering from the University of Porto, Porto, Portugal, in 1995 and 2005, respectively. She is an Assistant Professor with the Faculty of Engineering, University of Porto, and a researcher at Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC). She has been involved in several large national and European R&D projects in areas such as decision support systems, transport, and mobility. Her main research interests include combinatorial optimization, human–computer interaction, and transportation.

João Falcão e Cunha received the M.Eng. in electrical engineering from the University of Porto, Porto, Portugal, in 1983; the M.Sc. in operations research from Cranfield University, Bedford, U.K., in 1984; and the Ph.D. in computing science from Imperial College London, London, U.K., in 1989. He is the Dean of the Faculty of Engineering, University of Porto, and a researcher at Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC). He has been involved with theoretical and experimental work in software engineering and information systems applied to transportation systems for the past 25 years. His research interests include decision support systems, graphical user interfaces, object-oriented modeling, and service engineering and management. Dr. Falcão e Cunha joined the IEEE in 1982 and is now a Senior Member.