Paper submitted for publication in:
Conference on Advances Systems in Public Transport

Using Farecard Data for Schedule Matching and Evaluating Transit Service Performance

Siddharth Gupta, sid1.gupta1@gmail.com, +91-4422574250
Undergraduate student, Infrastructural Civil Engineering, Indian Institute of Technology-Madras, India
Occupational Trainee, University of Queensland, Australia

Mark Hickman, m.hickman1@uq.edu.au, +61-733653692
Professor and ASTRA Chair of Transport Engineering, University of Queensland, Australia

Submission Date: August 31, 2014

Abstract

In this paper, we present an approach for mapping trips as observed in the farecard data of a transit system to the scheduled trips as indicated in the General Transit Feed Specification (GTFS). The mapping process is more complicated than that in an Automatic Vehicle Location system since data are not always available at each stop on a trip. Also, the process is exposed to a much broader range of data and operation inefficiencies (due to possible negligence on part of passengers and drivers in addition to the system inefficiencies). A cogent mapping with farecard data can enable us to build information systems that explore a broad scope of transit service characteristics. Based on the proposed approach, the paper looks at ways of exploiting the richness of the data and overcoming shortcomings inherent to the data for providing service reliability and occupancy information, among others. We also discuss an approach for improving the schedules for transit services based on observed operations that can account for variations in travel time at different times of the day.

Keywords: Farecards, Schedule Matching, Transit Reliability, Transit Occupancy, Information Systems

1. Introduction

Public transport farecard data open up the possibility of finding a wide range of information regarding the performance of individual vehicle trips in a transit network. Many technologies have been implemented to monitor and manage operations, including Automatic Vehicle Location (AVL) systems, which track vehicle movements during operations, and Automatic Fare Collection (AFC) systems, which allow passengers to pay fares using automated, cash-free means. These fare payments, using farecards or similar technologies, allow one to identify passenger boarding locations from a tap-on while boarding, and in some cases alighting locations when a tap-off is necessary.

In order to gain information connecting service supply to passenger demand, we need to be able to map passenger trips as observed in the farecard data to the scheduled vehicle trips as indicated in the General Transit Feed Specification (GTFS) (Google, 2014). This task can be achieved using data from Automatic Vehicle Location (AVL) systems that often accompany implementations of Automatic Fare Collection (AFC) systems. Such analysis has been promoted in a variety of studies, including Bertini and El-Geneidy (2003) and Furth et al. (2006), where archived AVL data can be mined extensively to estimate vehicle travel times, schedule adherence, and other measures. For AFC systems, much of the interest to date has been oriented toward discovering passenger behaviour, based on the ability to identify origins, destinations, and time of travel (see Pelletier et al. (2011) for a recent summary). Some AFC data have been used for more aggregate measures of transit performance (Trépanier et al., 2009; Navick and Furth, 2002), while only very limited research to date has explored the exclusive use of AFC data for vehicle schedule adherence, travel time studies, and trajectory analysis (Chu and Chapleau, 2008; Lee et al., 2012; Sun et al., 2012).

Yet, in the absence of a strong complementary AVL system, we may be able to use the farecard data alone to be able to identify these vehicle trips and their mapping in the scheduled operations. This paper is focused on an approach to do so.

Once the schedule matching is performed, we can map all the characteristics derived from the farecard data onto the scheduled trips. This process can hence enable the provision of wider variety of travel information, including occupancy, service reliability, and passenger demographic composition, than is provided in the absence of AFC systems. Also, since the GTFS often does not account for variations in inter-stop travel times at different times of the day, we can use results from the mapping to propose modifications to the schedules so that they more closely match the actual operations.

This paper starts in the next section by exploring how individual trips can be identified from farecard data. In the third section, we explore the different means of mapping farecard (passenger) trips to GTFS scheduled vehicle trips and the various caveats associated with this process. Since, unlike AVL data, farecard data are not always available at each stop along the route, in the fourth section, we look into how we can use interpolation and extrapolation techniques to determine the missing arrival and departure times. In the fifth section, the paper finally discusses the difference in approach when it comes to providing information pertaining to transit vehicle

occupancy and the application of this process towards updating the transit vehicle schedule.

This paper uses data from the AFC system implemented in Brisbane's public transit network in the form of smart cards called Go Cards. The AFC system in Brisbane is a closed system, i.e. passengers are required to tap-on and tap-off during their journeys. The dataset is therefore able to provide information regarding boarding and alighting locations and associated timestamps. It also provides encrypted but consistent IDs of the cards used on the trips. The penetration rate of the farecards is 85-90% of all journeys in Brisbane, thus generating a nearly complete universe of journeys within the transit network.

2. Obtaining GTFS and farecard trips

Overview

Information regarding transit service operations for Brisbane's transit network is available in the form of Google's GTFS (Google, 2014). These files can be parsed to get detailed information regarding trip schedules on any route on any day within the span of the version of the GTFS files being used.

The mapping process is illustrated using the operations of a chosen route on a particular day. Since the process is easily generalized, it can be extended to any route for any chosen time period. The route, arbitrarily chosen for analysis, is route 150, and the operating date is 4 March 2013 (the first Monday of the month).

Routes often have different frequency and length of trips at different times of the day. A trip on a route can be defined completely by the sequence of stops on the trip along with the arrival and departure times at each of these stops. Hence, classes of trips are created based on the stop sequence that they follow. A particular stop sequence observed by a set of trips on a route is referred to as a 'pattern'. The GTFS for route 150 indicate the presence of 5 patterns. Three of these put together constitute the outbound trips on the route, and the remaining two account for the inbound trips. The combined Outbound and Inbound schedules for route 150 are shown in figures 1 and 2, respectively. The map of the route and the five constituting patterns are shown in figure 3.

The process of trip detection

We shall now focus on obtaining trips on route 150 as observed in the farecard data. Since farecard data are prone to a variety of systematic and non-systematic errors, we need to adopt a series of filters to ensure that these errors do not disrupt the authenticity of the detected trips. Prior to the implementation of the trip detection algorithm, we adopt 2 layers of filters; the first layer is common across all routes and the second is based on the route characteristics.

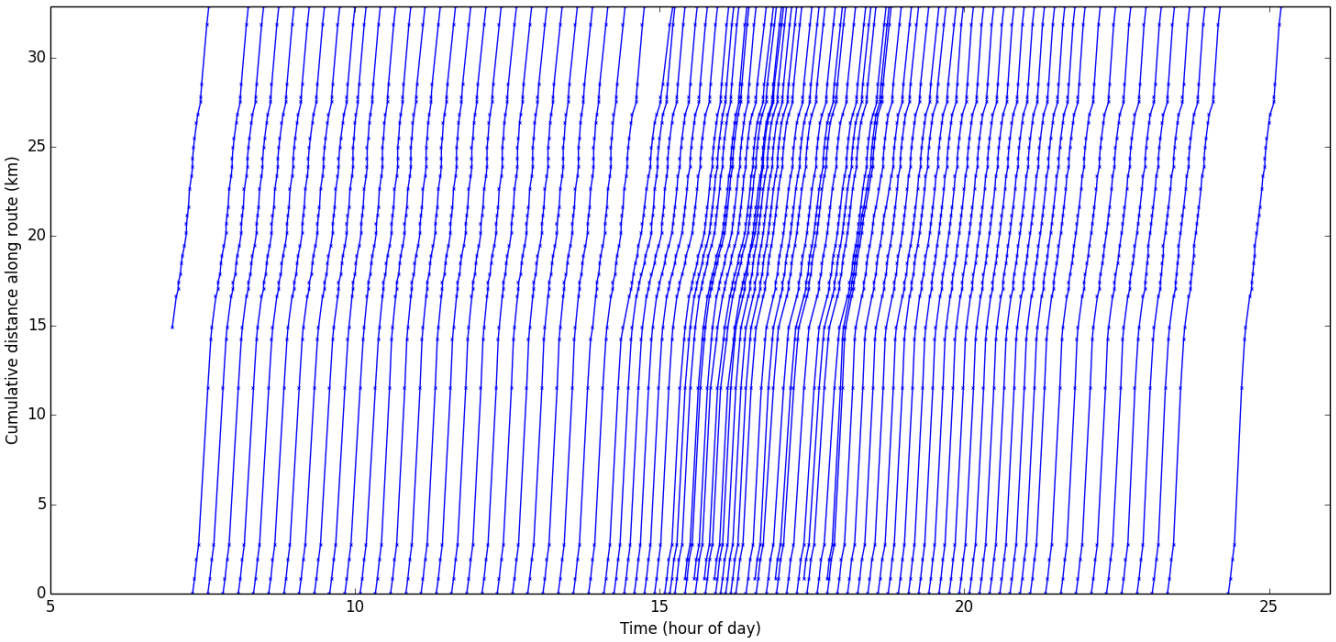


Fig.1: Outbound GTFS for route 150

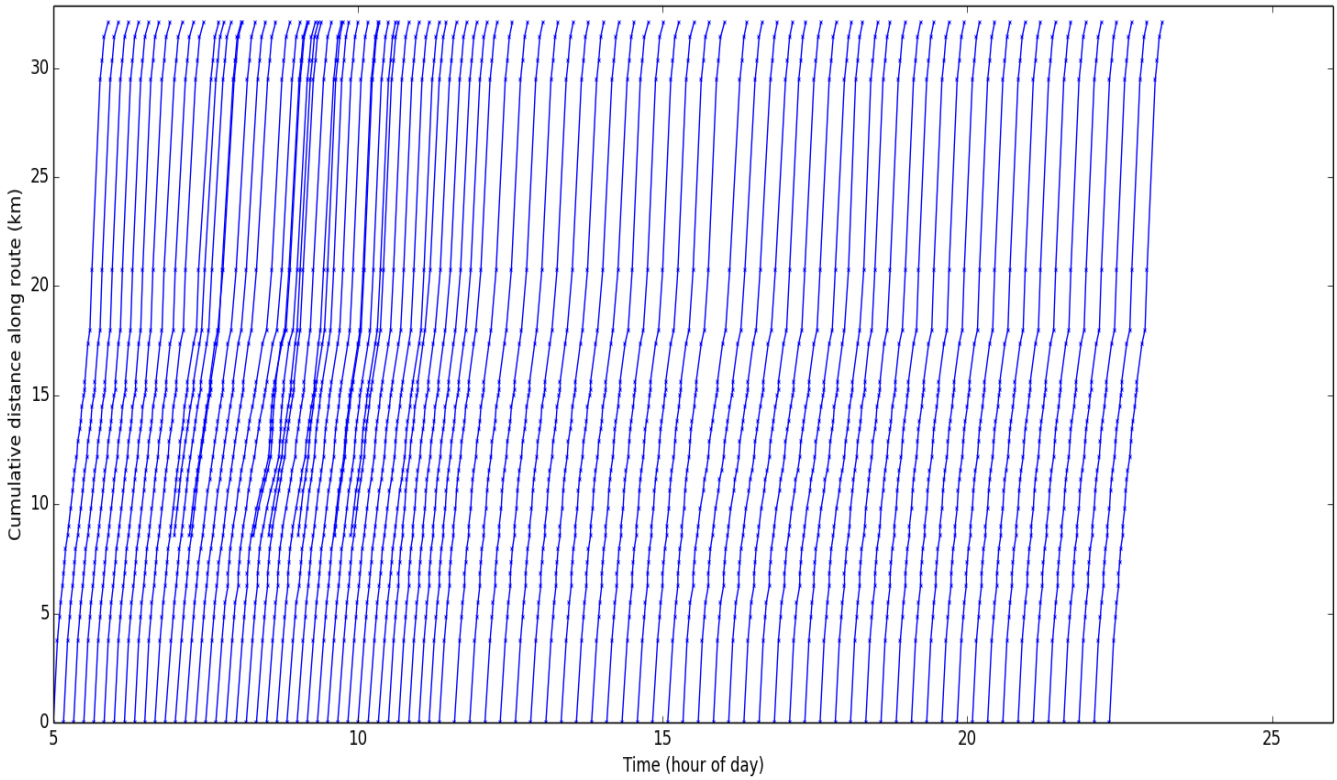
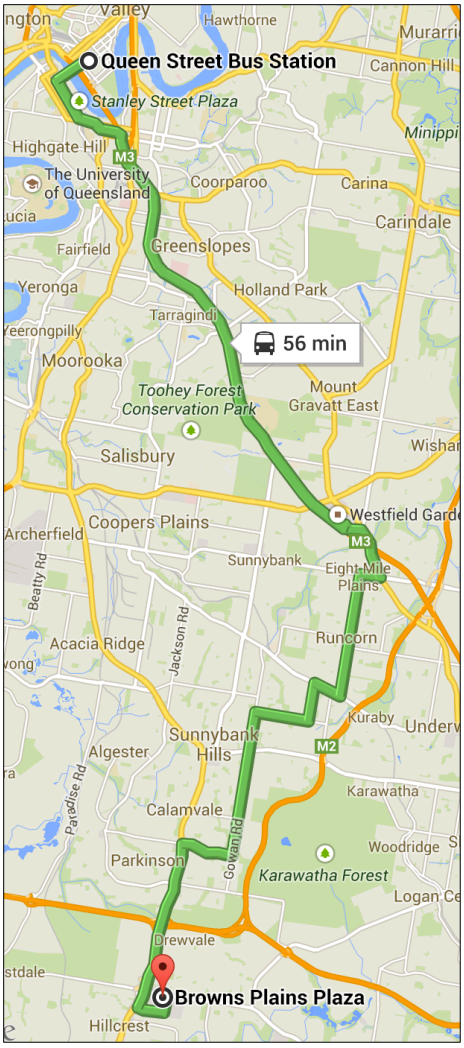


Fig.2: Inbound GTFS for route 150



Map showing route 150

Figures (a) to (e): The 5 trip patterns on route 150

Patterns (a), (c) and (e) together form the Outbound trips, while (b) and (d) together account for the Inbound trips

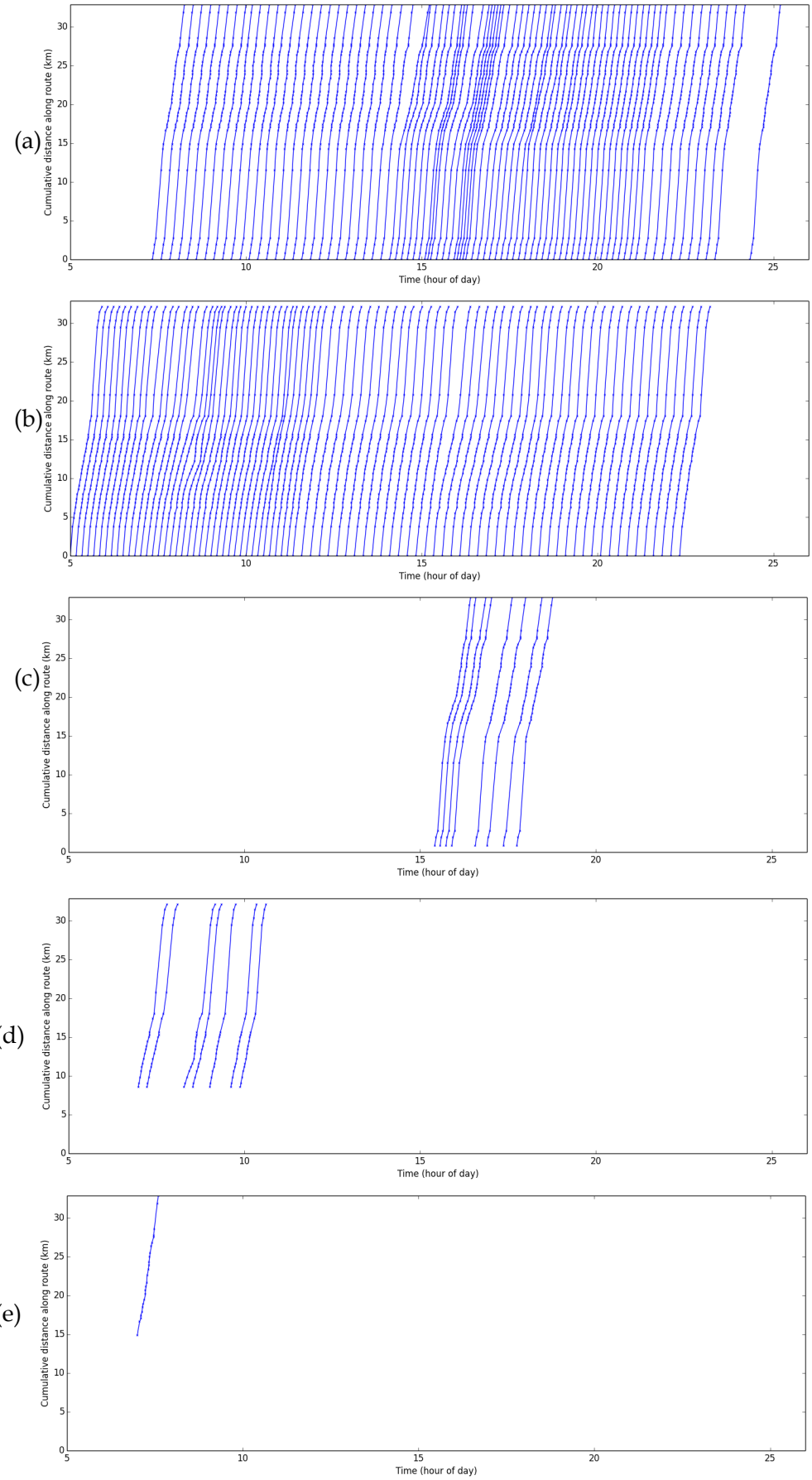


Fig.3: Map and patterns for route 150

The first filter layer ensures:

1. the observations do not have null fields
2. the boarding and alighting times are acceptable

The second layer composed of route-specific filters ensures:

1. the boarding and alighting stops in an observation are present in the same sequence on some pattern in the set of patterns for the route.
2. the traversal speed indicated by the observation should be between 10kmph and 100kmph. The distance used is that indicated by the shape files of the patterns.

After this, the observations for the route are segregated on the basis of the 'run' field, which is an indicator for individual transit vehicles. This is then fed into the trip detection algorithm described in the next section.

Based on the series of street segments served within the city, a transit vehicle plying on a route can serve either or both of the directions of a route at different times of the day. The farecard data contain a field indicating the direction of travel of the transit vehicle- a two-level direction flag assuming the value 'Inbound' or 'Outbound'. This value is recorded in the AFC system when the driver of the transit vehicle begins a route from a specific terminal. However, the driver may either forget to record this change, or may be delayed in doing so, leading to unreliability of the field for the purpose of analysis.

Hence, we need an algorithm to segregate individual trips performed by a transit vehicle on a route without falling back on the direction indicator. One such algorithm is described below.

Algorithm for trip segregation and determination of possible patterns for trip

The input to the trip segregation algorithm is the set of observations related to a transit vehicle on a route. The steps in the algorithm to segregate individual trips and to identify possible patterns is illustrated in figure 4 and explained below.

Stop sequencing

Boarding and alighting (tap-on and tap-off) activity of a passenger trip are separated and treated as independent events. All events at a stop that occur within a threshold time interval of each other are grouped together to represent a stop made by the transit vehicle at that location. The arrival and departure timestamps for the stop are computed as the least and the greatest timestamp of the observations grouped together. With these events at the stop level, the groups are arranged chronologically to get the sequence of stops made by the transit vehicle on the route.

Filtering the stop sequence

The alighting events recorded at the first stop in the stop sequence are disregarded, as these suggest passengers alight before any passengers board in that direction. The presence of the first stop-second stop pair is checked in all the patterns related to that route. If the pair is not present, the first stop is removed from the sequence; this trims

the stops to only include those observed in a given pattern. Once the first stop has been determined, for all subsequent stops on the run, this test is modified to consider if either of the previous stop-current stop or current stop-subsequent stop pairs is present in the pattern. If neither is present, the current stop is eliminated from the sequence. This allows unserved stops to be eliminated from the possible patterns.

If the stop pair is present, speed validation is performed based on the departure time at the first stop and the arrival time at the second stop. The speed and stop pair is accepted if the speed is between 10kmph and 100kmph. If this test is failed with both the preceding and succeeding stop, the stop being validated is eliminated from the sequence.

Trip determination

After the stop sequence has been determined, we move onto finding the vehicle trips included in the sequence and the patterns that the vehicle is likely to follow. For this, starting with the first pair of stops, we determine the set of patterns in which each pair occurs sequentially. After this, we determine the intersection of the possible patterns for each of the pairs, again in the same order as they occur in the stop sequencing. At each point in the sequence where the intersection of patterns across consecutive stops becomes a null set, we assume the start of a new vehicle trip. In this way, the vehicle trips performed on each pattern are determined.

Trip concatenation

If a transit vehicle waits at a particular stop (without regular boarding or alighting) for more than 5 minutes, then the trip on which this occurs would be fragmented in the AFC data and observed as two trips. Though the occurrence of such an event is highly unlikely in the Brisbane network, the algorithm enables concatenation of trips by the same transit vehicle on the same pattern within a 15-minute interval. With this final concatenation step, the determination of trips by each transit vehicle on the route is complete.

Aggregation

The vehicle trip results from all the transit vehicles are aggregated to get the results for the route for the day.

Deviation of farecard trip count from GTFS trip count

It should be noted that the number of vehicle trips determined by this process can be lesser or greater than the number of GTFS trips. The reasons for this can be:

1. Deviation of operations from the scheduled GTFS (i.e., the possibility of higher or lower frequency on the given day)
2. An absence of farecard transactions on a trip actually observed on the network; For example, on occasions when public transit is free or if passengers opt in for paper tickets rather than farecards

3. Transit vehicles on another route/service using an incorrect trip number and/or the route number.
4. Incorrect clustering of timestamps to constitute a possible trip; for example, under bus bunching.

Though most of such errors should be filtered out, some of them can persist on some occasions.

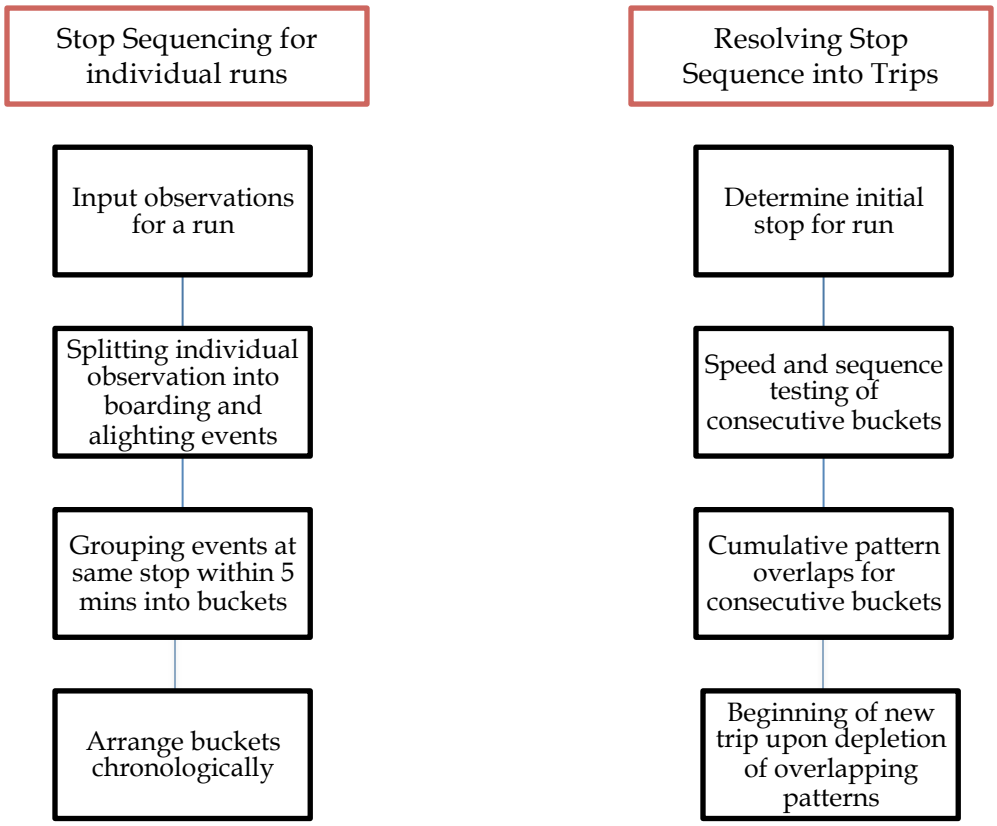


Fig.4: Logic used for determining the trips made by a transit vehicle on a route

The results of the vehicle trip determination and possible pattern mapping from the farecard data are shown in figure 5. It should be noted that at this stage, some of the trips can possibly be mapped to multiple patterns, since only the sequence of stops from the farecards has been considered. Once a temporal comparison with the GTFS trips is performed, farecard trips shall be resolved into unique patterns.

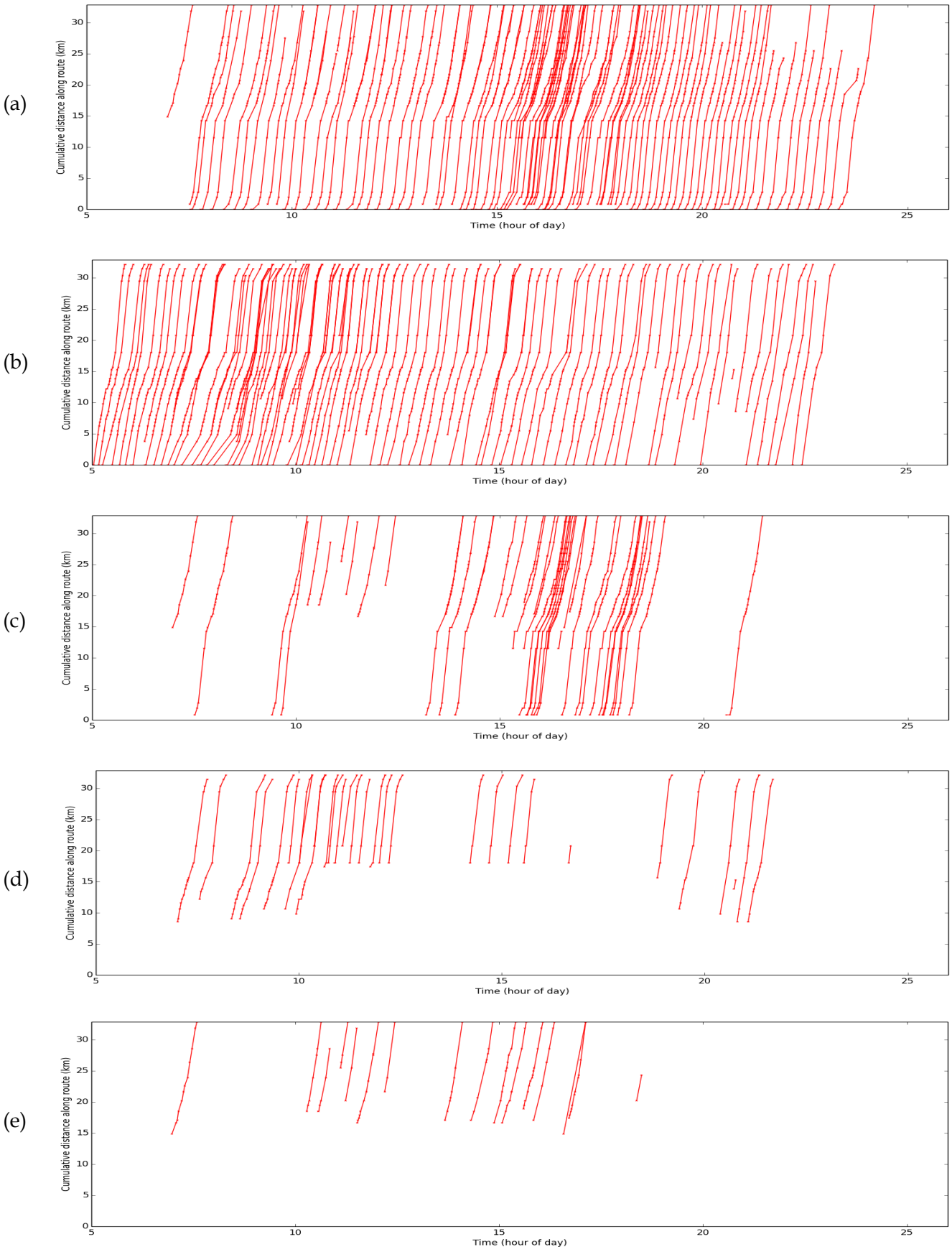


Fig.5: Trips determined from farecard data and the possible patterns with which they can match

3. Mapping farecard trips and GTFS trips

The process of mapping vehicle trips from the farecard data to the vehicle trips in the GTFS can be initiated in different ways. We can allow farecard trips to be mapped to the most similar GTFS trip or allow GTFS trips to select the farecard trip most similar to them. We can also choose the type of mapping that we want between the two sets: one-to-one or many-to-one. We should also make a provision for the absence of a mapping for some trips in both sets, for cases where a trip in one set is not observed in the other.

Since the number of trips in the two sets is usually found to be different (with both positive and negative deviations from the GTFS count), we have chosen a many-to-one mapping from the farecard trips to the GTFS trips. For the mapping from GTFS trips to farecard trips, we should opt for a one-to-one mapping since we want trip information to be based on the actual matching of a single GTFS trip.

Many-to-one mapping from the set of farecard trips to the set of GTFS trips

The difference between a farecard trip and a GTFS trip was measured in terms of total absolute deviation in arrival time and departure time at each stop along the route. The summation was taken across all stops for which information was available on a farecard trip. The lower the deviation, the higher is the similarity between trips.

The results of this mapping process are shown in figure 6. As one would expect, some of the GTFS trips have not been matched with any farecard trips. Since the combined plot is relatively cluttered, results for individual patterns have been illustrated. The unmapped GTFS trips for the inbound and outbound patterns are shown separately in figure 7. The time intervals when no mapping is found are likely characterized by bus bunching. Further analysis can be done to determine the cause and starting point of the bunching process. Other GTFS trips, such as the one operating close to midnight in pattern 1, might genuinely have no counterpart in the observed trips.

One-one mapping from the GTFS trips to the farecard trips

The total deviation of a farecard trip from a GTFS trip depends on the number of stops on the trip. If total absolute deviation were used for this comparison, the dissimilarity of the GTFS trip with short farecard trips having consistently high deviation at all stops would be underestimated, when compared to the dissimilarity with a longer farecard trip with low deviations at more stops. Therefore, we should opt for absolute deviation per stop as a deviation metric.

For cases where a farecard counterpart for a GTFS trip is nonexistent, we need to have an upper limit on the deviation metric. This should be based on the trip frequency during the scheduled period of the GTFS trip and can be set to half of the time interval between consecutive trips, per stop (hence the mapping is not truly a one-to-one mapping). Alternately, in case the farecard trips outnumber the GTFS trips, we shall be left with surplus farecard trips that can either be left over or mapped to the closest GTFS if the nature of the additional trip suggests so.

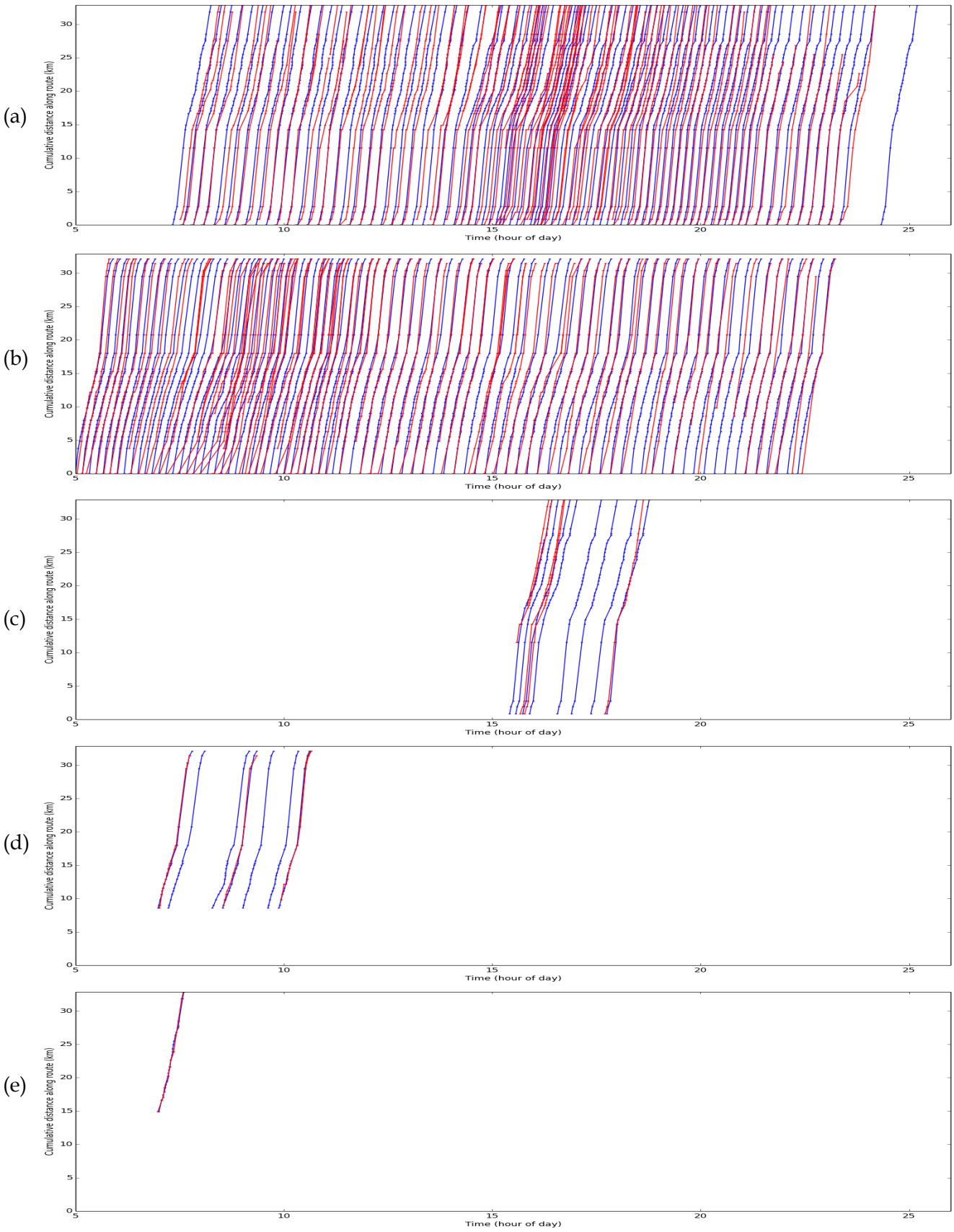


Fig.6: Trip matching from the farecard set to the GTFS set

(a) Inbound

(b) Outbound

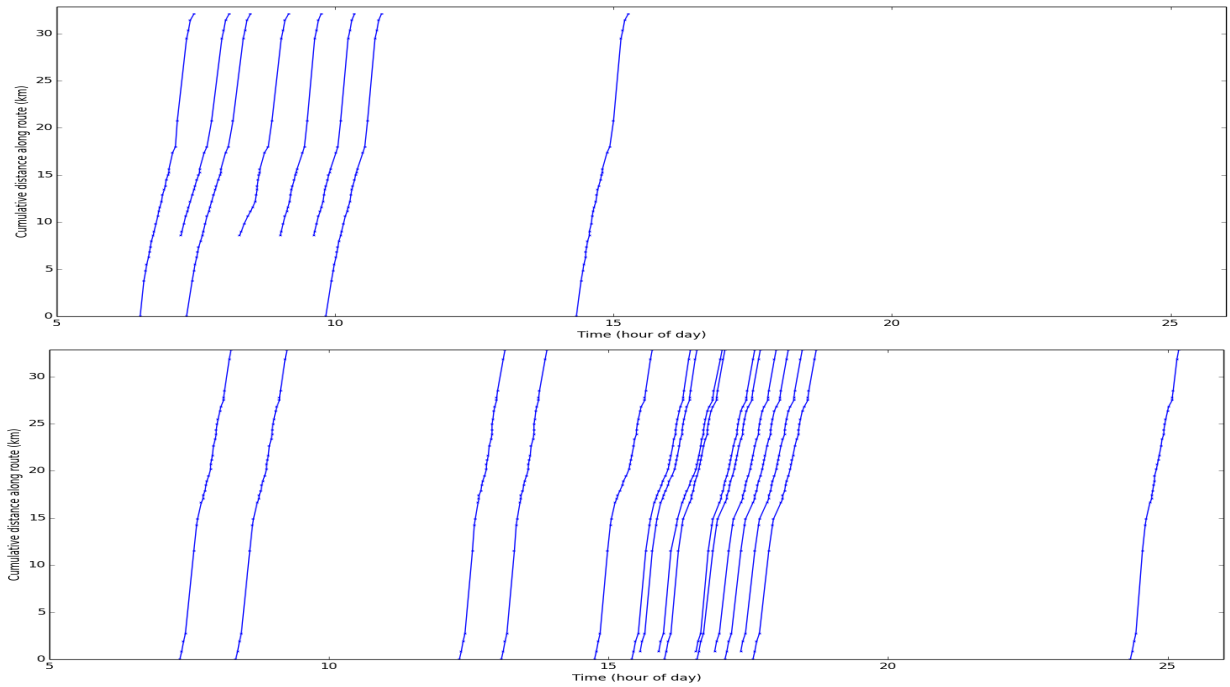


Fig.7: Unmapped GTFS trips from the many-to-one farecard to GTFS matching

Pattern preference with one-to-one mapping

As with the route under consideration, vehicle trips can follow multiple patterns on a route. A GTFS trip on a short pattern might have a unique possible mapping while a trip with a longer pattern in proximity of the shorter trip might possibly be matched with its own 'actual' counterpart, but also may be matched with the actual counterpart of the trip with the shorter pattern. If the trip with the longer pattern turns out to be more similar to the actual mapping than the trip with the shorter pattern, and first preference is given to the trip with the longer pattern, the trip with the shorter pattern might be left unmatched despite the existence of a counterpart in the farecard data. To avoid such errors, we can allow shorter patterns to be matched prior to the matching for longer patterns. In addition, the mapping here has been performed chronologically within a day.

The results of the one-to-one mapping process, with a maximum deviation of 7.5 min per stop, are shown in figure 8. The mapping process has been able to determine farecard counterparts to most of the GTFS trips.

4. Utilizing results from schedule matching

The process of schedule matching relies on the accuracy and availability of the boarding and alighting information, both in terms of location and in terms of the timestamp. Since the process can be achieved by using even a relatively small number of farecard observations from a vehicle trip, we were able to implement multiple filters to utilize only those observations that were likely to be accurate.

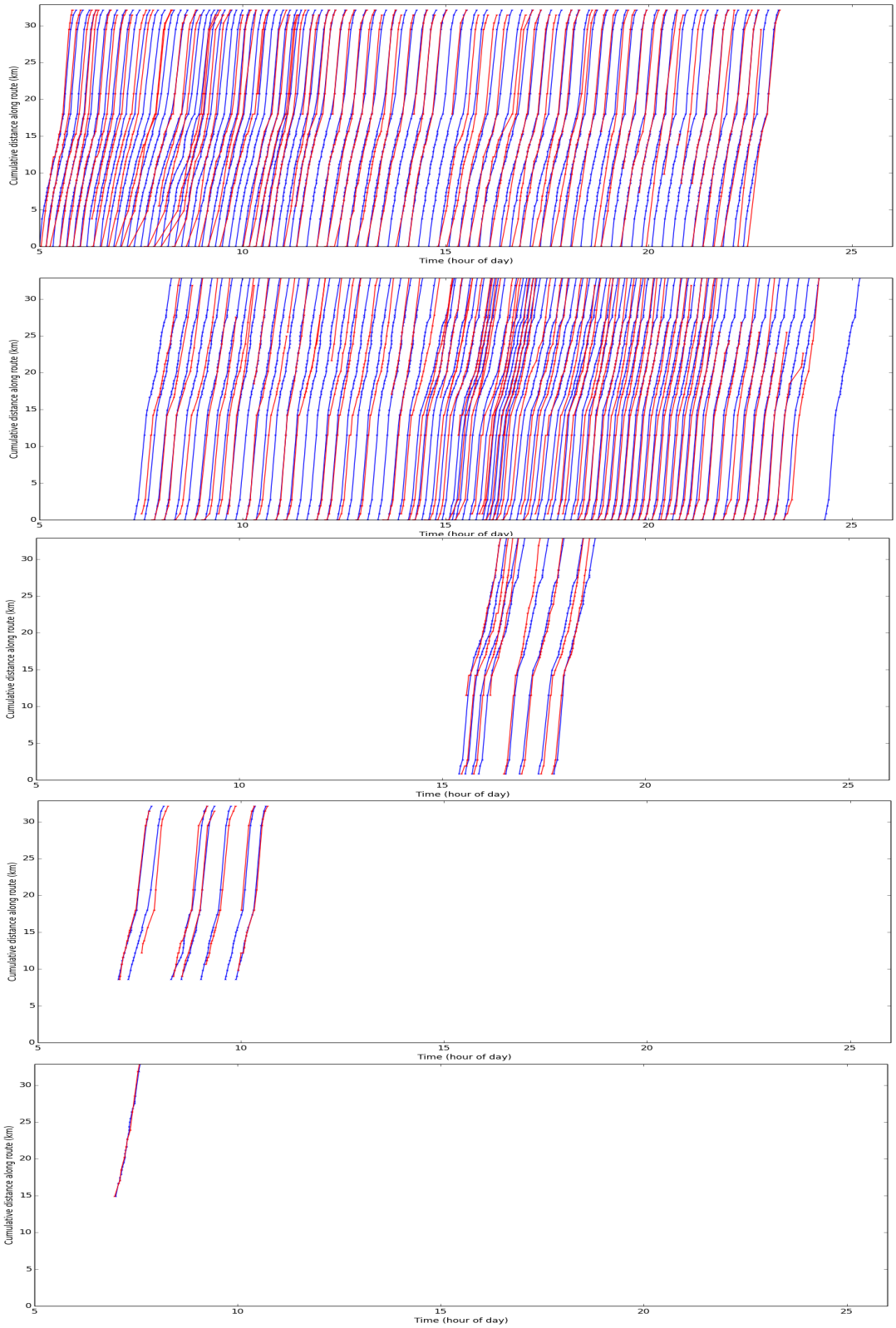


Fig.8: Results of one-one mapping from the set of GTFS trips to the set of farecard trips

The applications of schedule matching, though, might not rely heavily on the accuracy of the boarding and alighting locations and timestamps. For example, in the process of predicting the occupancy of transit vehicles, relatively fuzzy timestamps might be permissible so long as the boarding and alighting locations are accurate. Similarly, while analyzing the demographics of the commuters on a route, focus might shift towards ensuring the general validity of an observation without high accuracy caps on any individual field.

Once the process of matching farecard data to the GTFS vehicle trips is complete, we can retrace our steps to accumulate all the farecard transactions between the first and last observations of each full vehicle trip. This second pass at aggregating vehicle boarding and alighting behavior provides us with more a complete dataset to reliably provide broader information regarding characteristics of individual vehicle trips.

Calibrating the GTFS schedules based on observed operations

The schedules contained in GTFS often do not account for variations in travel times between stops at different times of the day. When the process of trip matching is extended over multiple days having the same GTFS schedule, we can explore consistent variations in travel time and suggest modifications to the schedule for improving reliability of services. With our example, Route 150 had the same GTFS schedule for all weekdays except Fridays and public holidays. For illustration, we consider trips starting in the period from 7pm to 10pm to illustrate consistent deviations from the travel times specified in the GTFS. The dates chosen for analysis were March 1 2013, to April 14, 2013. This period had the same GTFS schedule for route 150 Monday through Thursday each week except Monday, April 1, which was a public holiday.

The deviations in arrival time at each stop on different GTFS trips at different periods of time highlight the need to revise the vehicle schedule. This is shown in figure 9. The light green line joins the mean arrival time taken over all observations at each stop. The four plots sequentially explore the deviations in arrival times at each stop for trips starting at approximately 7pm, 8pm, 9pm and 9:30pm respectively. If we follow the green line, it can be inferred that an average trip at each of these times starts off ahead of schedule, loses ground during the middle section of the route (being behind schedule on these stops) and towards the end of the route gets ahead of schedule. This suggests that the inter-stop travel times in the timetable are underestimated at the beginning of the schedule, appropriately estimated in the middle of the route (since there is nearly constant delay at all intermediate stops), and overestimated towards the end of the schedule.

Since this pattern is served repeatedly at different times of day and over a one and a half month period, it is likely that the travel times are actually misrepresented in the given timetable The means indicated by the nodes of the green lines can be used as more reasonable substitutes for the current timetable arrival times.

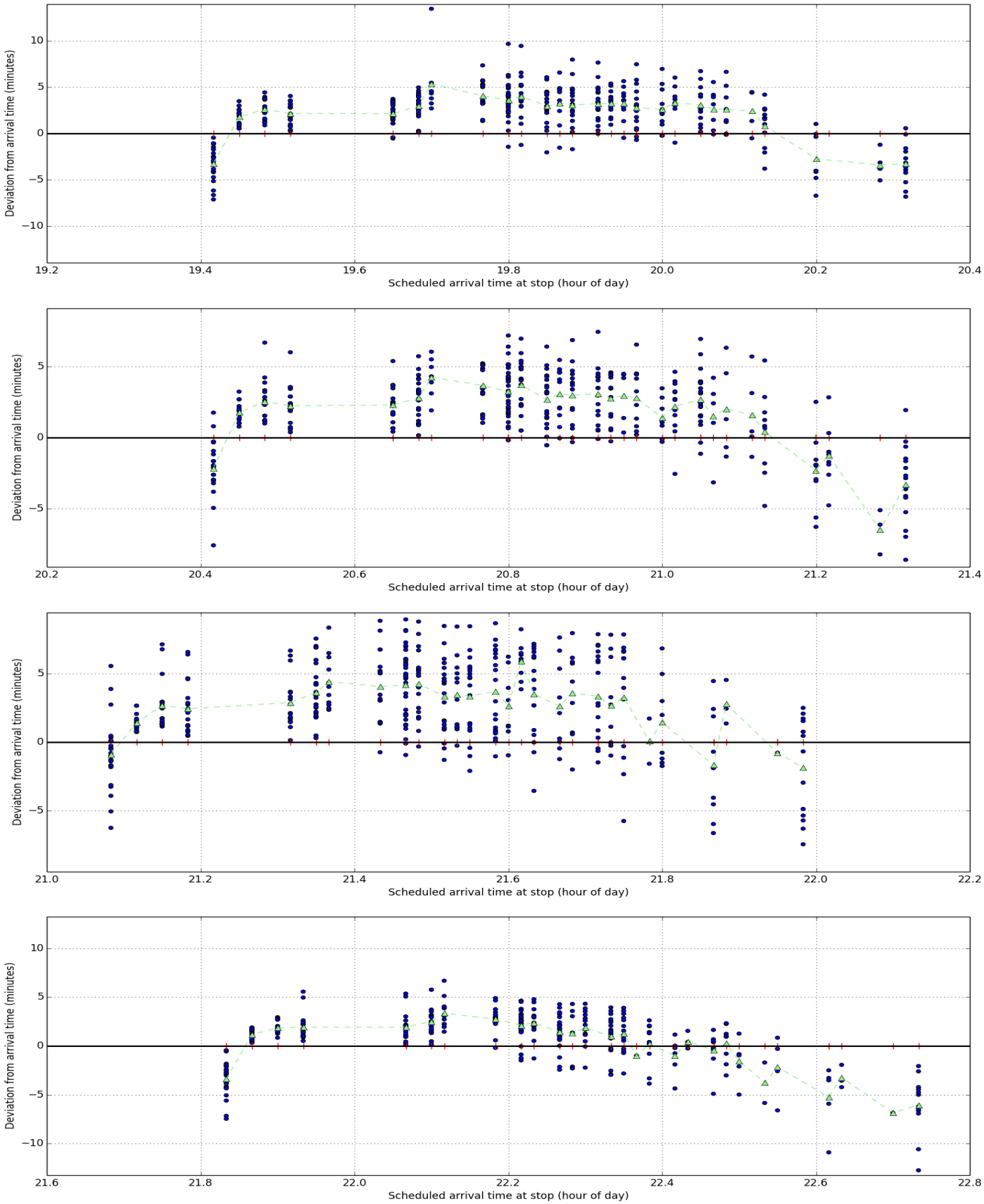


Fig.9: Exploring the possibility of GTFS calibration with observed trips from farecard data

Completing the actual arrival and departure time arrays for farecard trips

Vehicle trips taken from the farecard transactions depend upon the presence of boarding or alighting taps with accurate timestamps in order to determine the actual arrival and departure times at a stop. It is often possible that some stops on a trip do not observe boarding or alighting of passengers. In such cases, we need to come up with a technique to be able to interpolate or extrapolate the missing timestamps.

Missing timestamps at stops within the limits of an observed vehicle trip have defined upper and lower bounds. The task of interpolation can be completed using different approaches: (1) we can base it on the ratio in which the time interval is divided in the GTFS schedule; or, (2) we can also look at the ratio into which the time interval has been divided in historic iterations of the vehicle trip. Also, to capture possible travel incidents, we can learn from adjacent trips on the same day. Each of these three is a good alternative under different sets of conditions.

Extrapolation for determination of arrival and departure times at stops outside the observed limits of the trip, though possible, is prone to much greater deviations from reality, especially when we have to extrapolate over multiple stops. The task would be carried out best by learning from all three approaches mentioned above.

5. Advantages and limitations of schedule matching with farecard data

As mentioned earlier, vehicle location data are generally used to accomplish the task of schedule matching. These are susceptible to fewer sources of errors and provide reliable timestamps at all stops along the route. In addition, they can also be used to provide information regarding the speed of transit vehicles. In situations where transit vehicles and private vehicles share the right of way, these location data can even be extended to determine traffic flow characteristics and to identify real-time congestion- a task that is virtually impossible with farecard systems. At the same time, using the bounding timestamps from AVL systems for trip matching does not supersede the need for filtering outliers or performing the matching as detailed in section 3.

If we restrict our views to schedule matching alone, the only real advantage that the use of farecard systems might provide is the determination of the actual time interval at a stop that is utilized by passengers for boarding and alighting. This can enable us to adjust the arrival and departure times, if required, at each location.

Schedule matching with farecard data, however, also opens up several other opportunities that cannot be explored with vehicle location data. It opens the doors for providing much broader traveler information as already discussed. In addition, it can also be employed towards evaluating the utilization of transit services at different times of the day. This can be used as a basis for route truncation or elongation, as dictated by optimal utilization and can also help in exploring the possibility of altering service frequency. It can be extended further towards restructuring the stopping pattern on a route by quantifying the number of people likely to be affected by the removal or addition of intermediate stops.

6. Conclusions

Identifying vehicle trips from farecard data is a task that involves diligent cleaning and consideration of several operational intricacies that might be unique to individual transit systems. Once farecard trips are determined, different approaches for matching them to scheduled trips can help in capturing perspectives of both travelers and operators. In transit systems characterized with high deviations from scheduled operations, more detailed learning models can be generated to determine locations and causes of inefficiencies. Nonetheless, the implementation of a trip matching system can significantly assist in network planning, performance assessment and information provision.

References

Bertini, R., and A. El-Geneidy (2003). Generating transit performance measures with archived data. *Transportation Research Record 1841* , Transportation Research Board, Washington, DC, 109–119.

Chu, K., and R. Chapleau (2008). Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record 2063*, Transportation Research Board, Washington, DC, 63-72.

Furth, P., B. Hemily, T. Muller, and J. Strathman (2006). Transit Cooperative Research Program (TCRP) Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management. Transportation Research Board: Washington, DC.

Google (2014). Google Transit. https://developers.google.com/transit/, Accessed August 31, 2014.

Lee, D.-H., L. Sun, and A. Erath (2012). Study of bus service reliability in Singapore using fare card data. In *Proceedings of the 12th Asia-Pacific Intelligent Transportation Forum*.

Navick, D., and P. Furth (2002). Estimating passenger miles, origin-destination patterns, and loads with location-stamped farebox data. *Transportation Research Record 1799*, 107-113.

Pelletier, M.-P., M. Trépanier, and C. Morency (2011). Smart card data use in public transit: A literature review. *Transportation Research – Part C*, **19**(4), 557-568.

Sun, L., D.-H. Lee, A. Erath, and X. Huang (2012). Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. Proceedings of the ACM UrbComp '12 Conference, Beijing, August 12, 2012.

Trépanier, M., C. Morency, and B. Agard (2009). Calculation of transit performance measures using smartcard data. *Journal of Public Transportation*, **12**(1), 79-97.