CASPT 2015

Journey-Based Characterization of Multi-Modal Public Transportation Networks

Cecilia Viggiano · Haris N. Koutsopoulos · Nigel H.M. Wilson · John Attanucci

Abstract Planners must understand how public transportation systems are used in order to make strategic decisions. Smart card transaction data provides vast, detailed records of network usage. Combined with other automatically collected data sources, established inference methodologies can convert smart card transactions into complete linked journeys made by individuals in the public transit network. However, for large, multi-modal networks it can be challenging to summarize the journey records meaningfully. This paper develops a method for categorizing origin-destination (OD) pairs by the mode or combination of modes used. By aggregating across OD pairs, this categorization scheme summarizes the multi-modal aspects of network usage. The methodology can also be applied to subsets of data filtered by time of day or geography. The categorization results can inform performance analysis of OD pairs, allowing planners to make comparisons between pairs served by different combinations of modes. London Oyster card data is analyzed to illustrate how the OD pair categorization can characterize a network, allowing planners to quickly assess the roles of different modes, and perform OD pair analysis in a multi-modal network.

Keywords Multi-modal · Network Structure · Smart Card · User Behavior · Performance Evaluation · Journey-Based

1 Introduction

In public transportation, planning takes place at multiple levels. Pelletier et al (2011) suggest that smart card data can be used at three levels of planning: strategic, tactical, and operational. This paper focuses on using smart card

C. Viggiano
Massachusetts Institute of Technology
77 Massachussetts Avenue, Bldg 1-235
Cambridge, MA 02139 Tel.: +1-650-387-8672
E-mail: ceciliav@mit.edu

data to aid decision-making at the strategic level, by better understanding the roles of different modes. Do bus services feed rail or are they used as a stand-alone service? Do bus routes provide an alternative to rail for certain trips? This understanding informs the evaluation of the existing network and the identification of opportunities for improvement and expansion.

Several studies have sought to improve understanding of user behavior using smart card data, most commonly with an aim to improve marketing. Utsunomiya et al (2006) combined smart card data with personal information provided by users to analyze typical patterns for different user groups and develop profiles for users of specific stations. Morency et al (2007) analyzed the regularity of daily patterns and classified users based on their public transportation usage. Bagchi and White (2005) analyzed smart card churn to inform targeted campaigns to retain users. These studies can inform decisions about advertizing, promotions, and fare structure.

In contrast, research on network structure has focused primarily on network topology. Garrison and Marble (1964) used measures derived from graph theory to assess connectivity of transportation networks. Other studies built on these measures to characterize network shape (Gordon, 1974), inform network design (Vuchic and Musso, 1991) and identify properties of network structure (Derrible and Kennedy, 2010).

Instead of beginning with network topology, this paper introduces a journey-based approach to explore the multi-modal properties of networks. Extensive usage of smart cards for payment makes this approach possible. Smart cards record passengers' entries and in some cases exits from the public transportation system. Several researchers have developed methodologies to reconstruct individuals' itineraries, inferring origins, destinations, and transfers (Chu and Chapleau 2008; Gordon et al 2013). In networks where smart card usage is prevalent, these itineraries can provide a comprehensive picture of public transit travel.

In large networks such as London, where there are typically upwards of 15 million smart card transactions per day, the data must be aggregated to provide insight. This paper proposes a methodology that consists of two steps: the clustering of stops and stations for grouping journeys into zonal OD pairs, and the categorization of OD pairs based on the mode or combination of modes used (in the context of this research, bus and rail). The categories can be applied to all OD pairs, a subset of OD pairs, or a subset of the journeys for each OD pair, allowing for exploration of temporal and spatial variation. This methodology provides a concise representation of the modal attributes of network usage that can inform planners making decisions about network structure.

The stop and station clustering and OD pair categorization processes can also serve as a foundation for planners seeking to evaluate performance and make comparisons across a multi-modal network. While performance indicators are often calculated at the route or line level, performance at the OD pair level more closely reflects passenger experience and can take into account multiple paths serving the same OD pair. The stop and station clustering

methodology defines zonal OD pairs that serve as the fundamental unit of analysis for OD level performance evaluation. Given that modes have different properties, planners may wish to take mode into account in assessments of of a multi-modal network. The categorization scheme proposed in this paper can be used to associate OD pairs with the set of modes used.

The methodology is demonstrated using the London public transportation network as a case study. Stops and stations are clustered into 1,000 clusters, and OD pairs are assigned to one of seven modal categories. Time of day and geographic variation in OD pair categorization is presented, and distance and speed profiles are estimated for the most populated categories.

2 Methodology

As inputs, the methodology requires data for a set of one-way complete journeys consisting of, at a minimum, initial and final stops or stations and mode or combination of modes used. The stops and stations are clustered based on location to assign journeys to zonal OD pairs according to their initial and final stops or stations. Then OD pairs are categorized based on the share of journeys by each mode or combination of modes.

2.1 Stop and Station Clustering

Smart card journeys are first grouped by their origins and destinations. The true origins and destinations for the journeys (such as the individual's home or office building) are unknown, but the first and last stop or station for the journey are taken as proxies. In many cases, individuals can select between multiple paths. Therefore, instead of treating individual stops and stations as origins and destinations, clusters of nearby stops and stations are used for analysis.

One way to group nearby stops and stations is to use existing zonal schemes, such as postcodes or census tracts. In these schemes, the zones are defined using roads as boundaries. Consequently, bus stops and rail stations tend to be at the borders of zones, increasing the likelihood that individuals are choosing between alternatives in two different zones.

Instead, we cluster stops and stations using the k-means algorithm. This algorithm assigns data points to clusters such that the sum of distances between the data points and their cluster's centroid is minimized (Lloyd, 1982). In this case, each data point is defined by the geographic coordinates of a bus stop or rail stations.

The k-means algorithm consist of three steps. First, a set of data points are selected as the initial centroids. The number of centroids corresponds to the user-specified number of clusters. Then, each stop or station is assigned to the closest centroid, measured using the euclidean distance between the coordinates. Once all points have been assigned, the centroids are recalculated

as the mean values of all points assigned to a given cluster. This process iterates until the locations of the centroids do not change significantly from one iteration to the next. The algorithm always converges, but may reach a local (instead of a global) minimum, specific to the the selection initial centroids. The k-means++ initialization was used to select these centroids. This initialization process ensures that the initial centroids are geographically distributed (not too close together), which has better results than completely random selection (Arthur and Vassilvitskii, 2007).

Given that all stops and stations in a network may not fit clearly into clusters, the resulting cluster membership is likely to vary depending on initial centroids. Instead, the algorithm results in one plausible grouping of nearby stops and stations. This grouping is useful compared to existing zonal structures, because instances of zonal boundaries that split closely adjacent stops and stations are reduced.

In the k-means algorithm, the number of clusters is a user-specified input to the k-means algorithm. One way to select the number of clusters is to use a score such as the silhouette score, which is a measure that evaluates cluster tightness (the closeness of points within a cluster) and cluster separation (distance between clusters) (Rousseeuw, 1987). However, if the data does not have a strong underlying structure of clusters, there may be little variation in this score. Given that the clusters in this case will be identified as starting and ending zones for journeys in the network, it is helpful to consider elements of the downstream analysis. If there are too many clusters, there will not be a significant number of journeys in each zonal OD pair and neighboring clusters may be so close that individuals consider stops and stations in multiple clusters. At the same time, the stops within each cluster should be in comfortable walking distance of one another in order to constitute valid alternatives. Too few clusters can result in walking distances that are unrealistic. The number of clusters can be adjusted to reflect different assumptions about access distance.

2.2 Categorization of Origin-Destination Pairs

The result of the stop and station clustering is a set of zones with each stop and station belonging to a single zone. Given data on complete one-way journeys taking place in the network, these journeys can be assigned to zonal OD pairs according to the zones of their initial and final stop or station. These complete journeys are also classified by mode, defined as either bus (all stages were by bus), rail (all stages were by rail) or combined (journeys including both bus and rail stages).

Next, OD pairs are categorized based on the share of journeys belonging to the three modal categories (bus, rail, and combined). Each OD pair is categorized as one of the following:

- primarily bus
- primarily rail
- primarily combined

– bus and rail
– bus and combined
– rail and combined
– bus, rail and combined

To assign pairs to categories, we define a dominance threshold and an existence threshold. The dominance threshold is the percentage of journeys by a given mode required to place the OD pair into the single-mode categories (primarily rail, primarily bus, and primarily combined). If the dominance threshold is 80%, OD pairs with 80% (or more) journeys by rail will fall into the primarily rail category, and likewise for primarily bus and primarily combined.

The existence threshold is the minimum percentage of journeys by a certain mode to include the mode in the category assignment. If the existence threshold is 10%, an OD pair with 5% rail, 65% bus and 30% combined would fall into the bus and combined category and not the bus, rail, and combined category. Figure 1 shows the seven categories with the modal percentages plotted to the left of the schematic showing paths between zones.

Planners may opt to use different thresholds. A very high dominance threshold will identify the OD pairs where users appear to be truly captive to a given mode. Conversely, if planners wish to identify OD pairs with similar modal splits the existence threshold can be raised and the dominance threshold lowered.

3 London Case Study

London provides an interesting example for this analysis because it has a multi-modal public transportation network and a high rate of population growth necessitating growth in the bus system. As planners make decisions how to accommodate increased demand for bus service, they must first understand the role of bus in the current network. Figure 2 shows a map of London's multi-modal network.

3.1 Oyster Data Set

The smart card data used for the case study consists of 14 days of Oyster (London's smart card) transactions. The data was processed using the methodology developed by Gordon et al (2013) which infers origin bus stops, alighting bus stops, and links stages of multi-stage journeys using automatic vehicle location data and geographic and time-based thresholds. For simplification, only journeys of up to three stages were included (with stages here defined as initiating with a smart card tap), which represent 99.4% of all journeys.

Figure 3 displays the distribution of journeys by mode in London for the data analyzed. Underground, Overground, and National Rail are grouped as rail for this analysis. In the Underground, passengers can transfer between lines without tapping their card during the transfer. Therefore, we do not

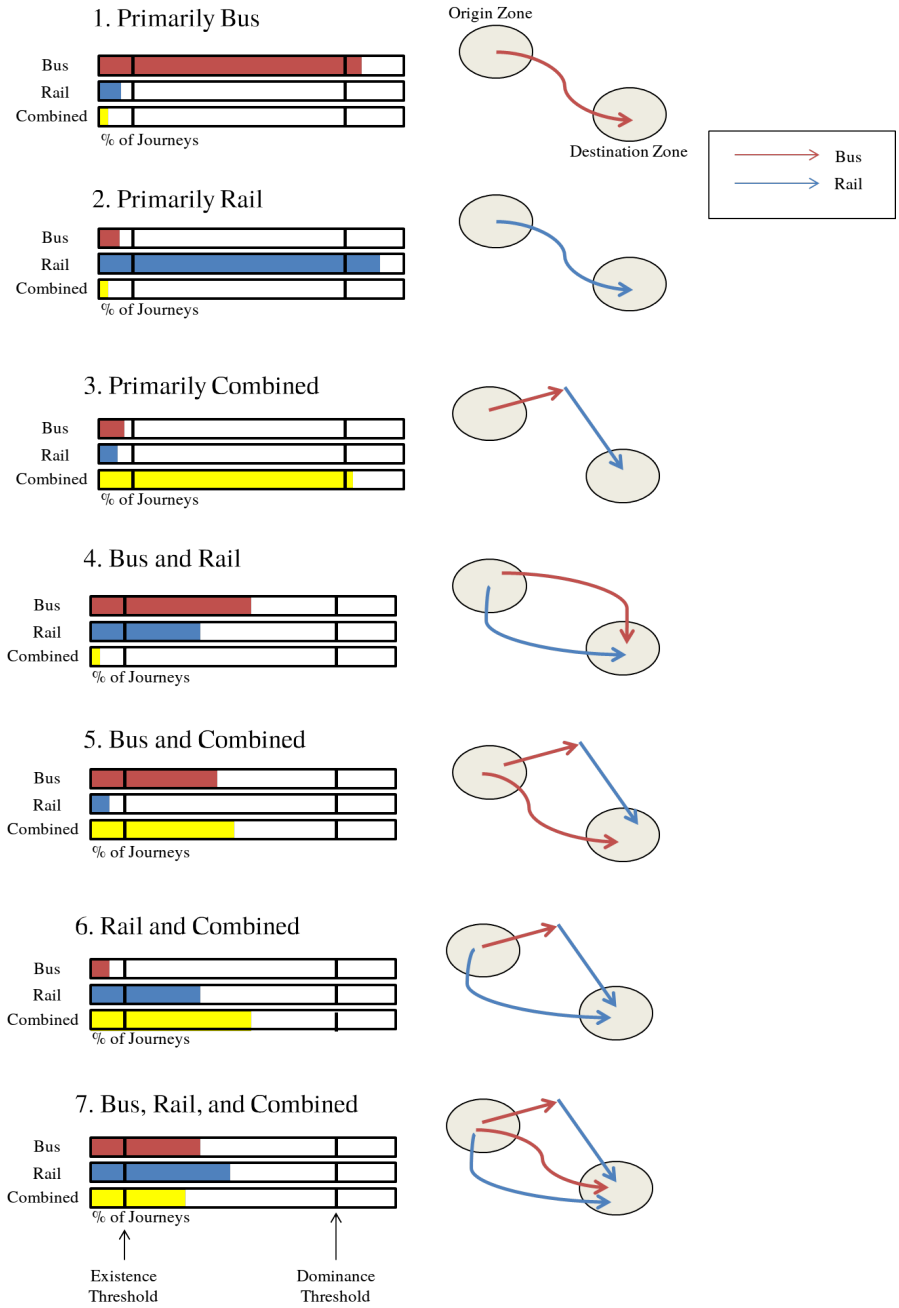

Fig. 1 Seven modal categories

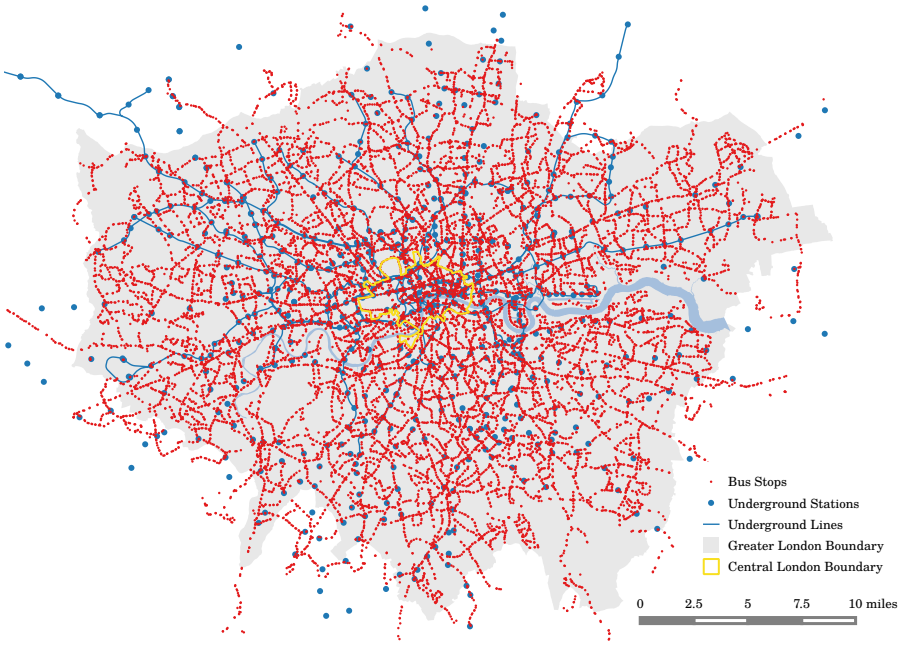

Fig. 2 Map of London rail stations and bus stops

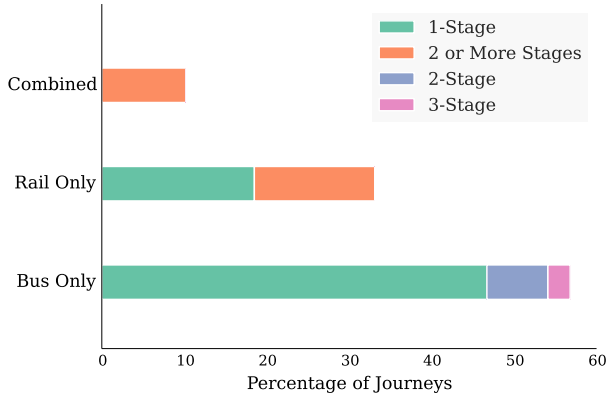

Fig. 3 Journeys by mode and number of stages

always know the exact path an individual takes through the rail network. If an individual's starting and ending station are served by the same line (i.e. the Victoria line), it is assumed to be a single rail stage. Otherwise the number of stages is designated as two or more. Passengers must tap their Oyster card at the beginning of each bus stage so these bus journeys can be accurately defined as one, two, or three stages.

The ODX methodology inferred the starting and ending stop or station for 81% of journeys made using an Oyster card in the two weeks analyzed. Of the Oyster journeys that did not have a starting and ending stop or station inferred, 69% were identified as single stage bus journeys with an origin stop and bus route but no destination stop inferred. This is because for bus stages, the ODX methodology infers the alighting stop based on the next stage of the journey or the same day return journey, meaning that destination stops for bus stages without a continuation or return journey cannot be inferred.

Because the proposed analysis is OD-based, journeys without destination stops cannot be included, meaning that the disproportionate inference rate for single stage bus journeys would result in under-representation of bus journeys. To correct for this under-counting, each of the single stage bus journeys with an uninferred destination stop was assigned an alighting stop by the following methodology: For each boarding stop, a destination stop distribution was constructed, consisting of the frequency of occurrence of all inferred downstream destination stops for single stage bus journeys originating at that stop. Then, for each journey beginning at that boarding stop which did not have an inferred destination, a destination stop was selected at random from this distribution. This methodology assumes that single stage journeys with uninferred destinations have the same destination distribution as single stage journeys with inferred destinations.

Through the inference methodologies, the Oyster data was transformed into a set of 90,306,224 complete journeys. Assuming an 80% Oyster card penetration rate (Transport for London, 2012) this accounts for approximately 75% of all journeys in the period.

3.2 Clustering Results

The k-means algorithm was applied to the the stops and stations in Greater London, with 1,000 clusters specified. The silhouette score revealed little difference between various numbers of clusters that would be consistent with a reasonable walking distance (800 to 1400 clusters). 1,000 clusters results in zones that average 1.6 km^2, but because stops and stations are more heavily concentrated in Central London, zones are smaller at the center and larger at the periphery. The number of stops and stations per zone varies as well, as shown in Figure 4. Figure 5 displays the zones generated for a portion of Central London. Due to their higher ridership, rail stations were weighted tenfold to increase the likelihood of their being close to the center of a cluster. When journeys are assigned to clusters, as expected, the zonal OD matrix is sparse; 48% of the OD pairs are empty.

3.3 Origin-Destination Pair Categorization Results

OD pairs were assigned to the seven categories outlined in Section 2.2, with an existence threshold of 10% and a dominance threshold of 80%. Only OD pairs

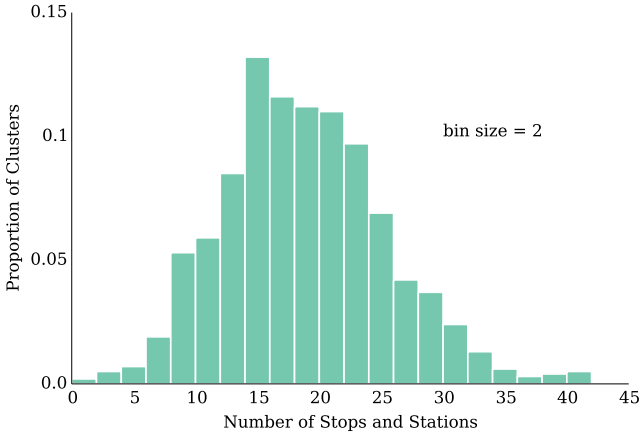

Fig. 4 Distribution of rail stations and bus stops per zone

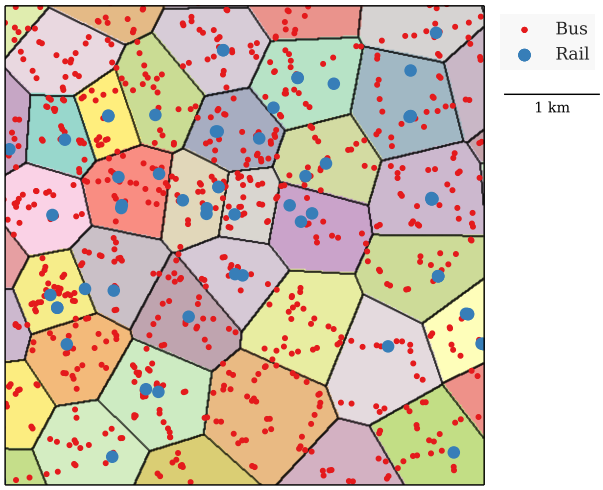

Fig. 5 Central London stop and station clusters.

with at least 385 journeys were included to avoid small sample size problems. Figure 6 shows the results. In 46% of OD pairs, bus is the primary mode. This may reflect the fact that many parts of the outer London network are served only by bus. Alternatively, users may have a strong preference for bus for certain journeys, for example short journeys.

Applying the methodology to subsets of the data can provide more detailed insight. Figure 7 shows the results of the methodology applied to weekday AM Peak journeys. This analysis can help planners understand how network usage

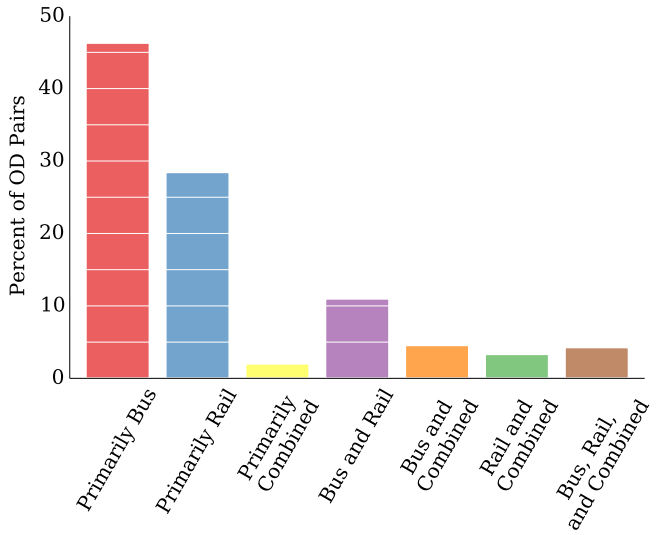

Fig. 6 Categorization of all OD pairs

changes over the course of the day. In London, the percentage of primarily rail OD pairs is greater in the AM peak than overall.

One can also consider geographic variation. To illustrate this, a central zone was defined and the following set of journey types were analyzed: OD pairs within the central zone (central), OD pairs that start outside the central zone and end inside it (to center), OD pairs that start inside the central zone and end outside it (from center), and OD pairs that start and end outside the central zone (periphery). Figure 8 shows how the categorization results differ between these journey types in the AM peak.

Central OD pairs include a high percentage of bus and rail designated pairs, indicating that both modes are important, in many cases providing parallel service. There is considerable asymmetry in journeys to and from the central zone. The primarily rail category dominates the "to center" OD pairs suggesting that rail is critical for these journeys. However, for the "from center" OD pairs, primarily bus dominates. This is likely due to the fact that destinations outside Central London are less likely to be close to rail stations. Even in the AM peak period, bus service is important for these reverse commuting trips. Not surprisingly, the bus network serves the peripheral trips almost exclusively.

3.4 Origin-Destination Pair Characteristics and Performance by Category

Planners can also use the categorization results to assess attributes of OD pairs falling into each category. These can be descriptive characteristics, such as the distance distribution or evaluation metrics, such as travel speed.

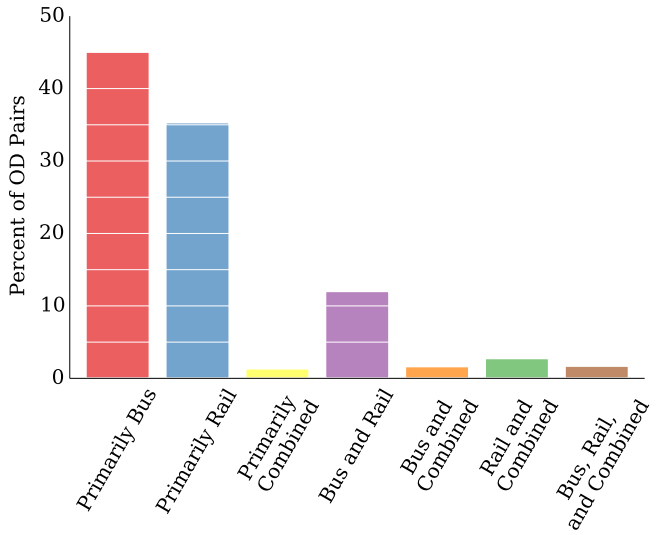

Fig. 7 Categorization of OD pairs for weekday AM peak

Figure 9 shows the distance distributions for OD pairs that are primarily bus, primarily rail, and bus and rail for the AM peak. Other categories were excluded from the figure because they make up a small percentage of AM peak OD pairs. Distance is defined as the straight line distance from the centroid of the origin zone to the centroid of the destination zone. This shows that in London rail tends to serve longer journeys from 2 to 8 miles while bus serves shorter journeys that range from 0 to 3 miles. The bus and rail OD pairs, in which some journeys are made by bus and others by rail, have a distance distribution falling between that of primarily rail and primarily bus, though it more closely mirrors the primarily bus distribution.

Figure 10 plots journey distance against journey time for a random sample of OD pairs, demonstrating the variation in speed across OD pairs. Again, only primarily bus, primarily rail, and bus and rail are shown because they are the dominant categories in the AM peak. Primarily bus OD pairs tend to have the slowest speeds but also have less variation in speed compared to primarily rail OD pairs. For OD pair level evaluation, planners can set different standards depending on the category that a particular OD pair is part of.

4 Conclusions

This paper presents a methodology for aggregating large quantities of smart card data in a meaningful way to help understand the ways in which passengers use different modes in a multi-modal public transportation network.

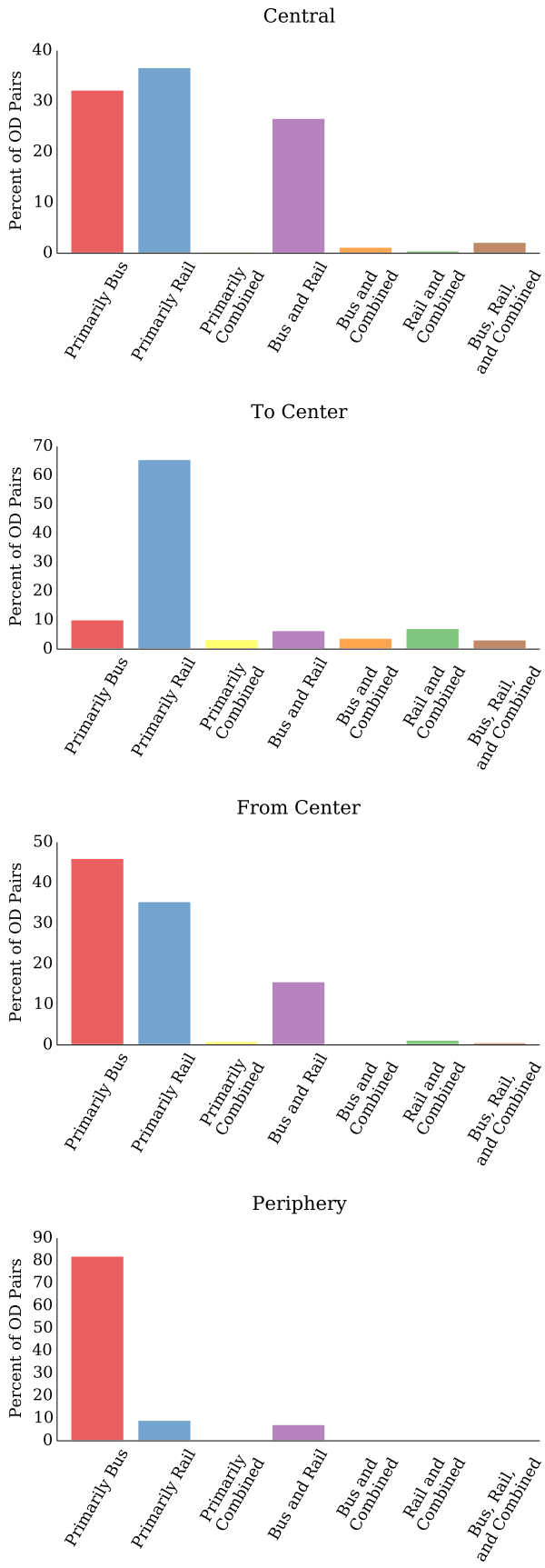

Fig. 8 Geographic variation in categorization results

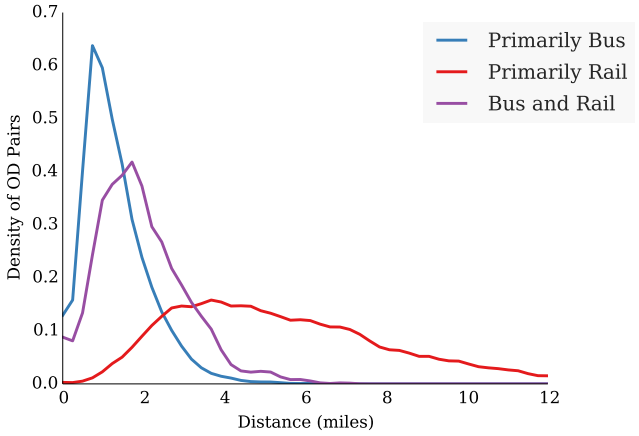

Fig. 9 Distance distributions of OD pairs by category

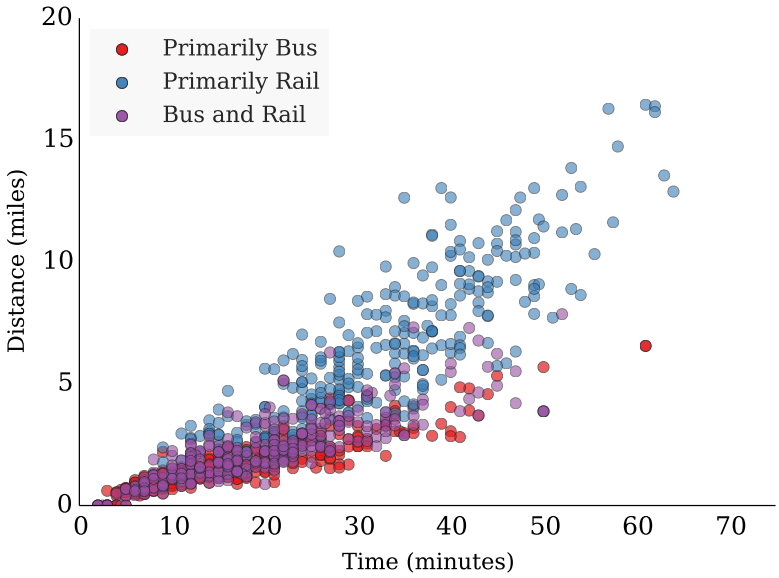

Fig. 10 Journey speed by category

Rather than making this characterization based on the network topology, this methodology starts from the passengers' journeys, and presents a clear picture of how the network is used, rather than of the services provided. This understanding of user behavior can inform planners as they make strategic decisions about network structure and modal expansion.

The analysis presented for the London network suggests that bus provides an important role for journeys around the periphery, journeys from the central zone to the periphery, and journeys within the central zone. With this information, planners can decide if they want to focus on improving the services that already fill these roles or concentrate on expanding the role of bus in journeys from the periphery to the center.

The categorization methodology provides a foundation for further evaluation of multi-modal public transportation networks. Performance metrics assessed at the OD pair level can be useful, particularly in a dense network where passengers may take different paths between an OD pair. Planners can opt to use category-specific standards for metrics to account for performance differences intrinsic to different modes.

Acknowledgements Many thanks to Transport for London for the support, guidance, and oversight of this research.

References

Arthur D, Vassilvitskii S (2007) k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms pp 1027–1035

Bagchi M, White P (2005) The potential of public transport smart card data. Transport Policy 12(5):464–474

Chu AKK, Chapleau R (2008) Enriching archived smart card transaction data for transit demand modeling. Transportation Research Record: Journal of the Transportation Research Board 2063(1):63–72

Derrible S, Kennedy C (2010) Characterizing metro networks: state, form, and structure. Transportation 37(2):275–297

Garrison WL, Marble DF (1964) Factor-analytic study of the connectivity of a transportation network. Papers in Regional Science 12(1):231–238

Gordon JB, Koutsopoulos HN, Wilson NH, Attanucci JP (2013) Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. Transportation Research Record: Journal of the Transportation Research Board 2343(1):17–24

Gordon SR (1974) Relationships between economies of scale and the shape of transportation networks. PhD thesis, Massachusetts Institute of Technology

Lloyd S (1982) Least squares quantization in pcm. Information Theory, IEEE Transactions on 28(2):129–137

Morency C, Trepanier M, Agard B (2007) Measuring transit use variability with smart-card data. Transport Policy 14(3):193–203

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

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20:53–65

Transport for London (2012) Join in the celebration across the capital this summer with a limited edition summer oyster card. Press Release

Utsunomiya M, Attanucci J, Wilson N (2006) Potential uses of transit smart card registration and transaction data to improve transit planning. Transportation Research Record: Journal of the Transportation Research Board 1971(1):119–126

Vuchic V, Musso A (1991) Theory and practice of metro network design. Public transport international 40(3/91)