Transit Ridership: Methods in Urban Studies and Transit Planning

Itto Kornecki

A Semester Project for the Institute of Terrestrial Ecosystems,
ETH Zurich

January 2019

Table of Contents

Abstract 3

1. Introduction 4

    1.1 Related Work 4

2. Dataset 6

3. Methodology 9

    3.1 Station accessibility 9

    3.2 Extracting commuters 10

4. Results 12

    4.1 Characteristics of commuting transit 12

    4.2 Transit and urban structure 13

    4.3 Accessibility analysis 16

    4.4 Hotspot analysis 18

5. Conclusions 22

    5.1 Future Work 22

References 24

Abstract

With the advent of Automatic Fare Collection in transit networks, there has been a dramatic rise in the availability of transit data. In this work, we show that this data can be used to reveal information about the urban structure and the accessibility of the rail network. Specifically, we analyze the diurnal ridership of the Transit for London underground stations in order to describe the urban structure of London and to quantify the suitability of the network for daily commuters. By removing the conventional dependence on origin-destination trajectories, our methods can be expanded to other transit networks, such as bus and tram. Our work serves as an easily applicable planning tool for urban planners and transit agencies.

1. Introduction

Over the past decades, Automatic Fare Collection (AFC) has been introduced into transit networks worldwide. Typically based on a smart card system, AFC allows passengers to digitally pay for rides on the transit network. As a side effect, AFC enables an unprecedented amount of transit ridership and mobility data. In the last decade, this data has become increasingly used in transportation research [1]. Nonetheless, AFC data is limited in that it is usually entry-only, lacking any information on where passengers exit. Therefore, the ridership data does not contain any origin-destination information. In addition, though AFC data has become increasingly common in transit research, it has not yet become a common tool in urban studies. In this work, we introduce novel methods to analyze ridership data both in transit planning and in urban studies. Moreover, our methods use entry-only ridership data, without the need for any origin-destination information, and can therefore be easily extended to other networks and cities.

1.1 Related Work

Numerous researchers have explored population density and population distribution as predictors of ridership, often introducing other predictors, such as walking distance or vehicle ownership [2][3]. However, whereas many researchers have attempted to predict ridership based on population characteristics, fewer have attempted to work in the opposite direction: estimating population and urban structure based on transit ridership. This is largely due to the availability of data: detailed populations censuses and geographic information system (GIS) data have been available for decades, whereas detailed ridership information has only recently become widely available with the advent of smart card systems. Nonetheless, despite its recency, several works have used AFC data to characterize urban structure. These works rely on a detailed record of passenger origin-destination trajectories for trips made on the network [4][5][6]. As such, these works can be grouped with similar works which use other origin-destination mobility data, such as mobile phone localization, in order to describe urban structure [7][8].

Although these methods produce substantial results, they rely on detailed passenger flows which are often impossible to obtain in other transit systems, such as tram and bus rapid transit (BRT) systems, where entrances to stations are recorded but exits usually are not. One approach to circumvent this is to predict origin-destination trajectories based on ridership data [9][10]. However, we show that the reliance on origin-destination information is misguided, and that a description of the urban structure can be made with only information of the entrances into the transit stations. Specifically, by analyzing the station entries over the course of a day, we are able to extract information about the workbound and homebound behaviour of the ridership, and therefore we can describe the home and work areas in the urban environment. This serves as a preliminary proof that information on the urban environment is possible even on transit systems where trajectory data is lacking.

In addition to a preliminary study of urban structure, we develop a novel method to explore the accessibility of stations. Transit accessibility is an extremely active field in recent years. Usually, researchers approach transit accessibility from a sociological perspective, attempting to find bias in the transit planning and service, and examining its effect on socioeconomic measures such as employment [11][12]. The main challenge in this research is to identify transit gaps: urban areas where public transit access is inadequate in some way, such as due to large walking distances from stations, inconvenient scheduling, or poor connections to desired stations. The usual method of finding transit gaps involves developing two indices: one to represent the transit needs and another to represent the transit service [13][14]. Demand is usually modeled based on age, income, and other socioeconomic factors such as vehicle ownership, while supply is modeled based on transit service properties, such as frequency and distance to the stations [14][15]. By subtracting the supply from demand, one is able to measure the adequacy of the service.

This method of estimating and subtracting demand and supply is heavily model-based, relying on several parameters. Existing methods therefore must take on the difficult task of estimating these parameters. We propose a novel method which sidesteps this issue by examining the difference between expected ridership and the actual ridership. We reduce the problem to two relatively accessible datasets: the population census and the transit ridership. By relying on more accessible data and fewer parameters, our method serves as an easily implemented transit planning tool which can readily be extended to other transit networks worldwide.

2. Dataset

Ridership

Our core analysis is done on the London Underground Passenger Count dataset, which is provided freely by Transit for London [16]. The dataset describes the average number of entrances and exits from each station in the Underground Network, represented as a time series spanning 24 hours. The time series is aggregated at 15 minute intervals, resulting in 96 data points per station. This represents an average of all days in the month of November 2017, separated into weekdays and weekends. We note that several stations which are part of the underground network are missing from the dataset and are therefore excluded from our analysis.

In addition to the Underground Passenger Count dataset, we also make use of the Rolling Origin-Destination Survey (RODS), which is provided freely by Transit for London [17]. The RODS dataset describes the number of trips between all stations, provided as origin-destination pairs. The trips are for weekdays only and are according to the time of day. Like the Passenger Counts, the RODS is based on rides in November of 2017.

Though the Passenger Count and the RODS datasets are similar, the former excludes trips made on other networks sharing the station (e.g. National Rail), while the latter does not. Therefore most of our analysis utilizes the Passenger Count dataset, while the RODS dataset is used only when we require origin-destination information. Specifically, the RODS dataset is used in two instances: first, when we show the symmetrical nature of trips in order to establish their commuting nature, and second, when we examine the flow between home and work hotspots.

Census data

Part of our analysis involves comparing the ridership to the population within the zones surrounding the station. We therefore require high-resolution data of the London population. For this, we utilize the official 2017 Census Output Area estimates [18]. Output Areas are the smallest census designation in the UK, thereby providing us with the highest resolution population estimates possible (Figure 1). This high resolution allows us to aggregate the census areas in order to estimate the population within each station zone. Although the 2017 census is only an estimate, it provides a better approximation of the existing London population compared to the full census performed in 2010, which is by now outdated.

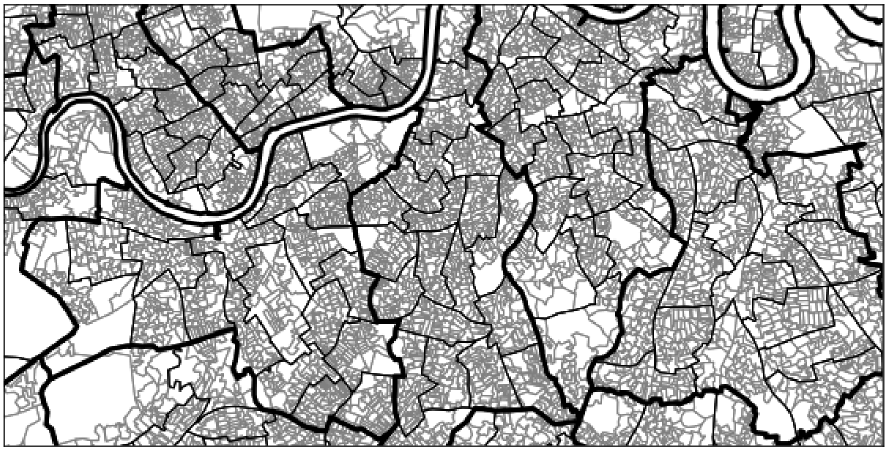

Figure 1: Portion of London showing Output Areas, wards, and boroughs.

The census population estimate is divided by age, allowing us to remove irrelevant age groups. Since we focus on home-to-work commutes, we restrict the census population to the working age population, which we designate as ages between 18 and 65. We do so in order to more accurately estimate the demand during morning and afternoon peaks. That being said, we note that there is a strong correlation between the working-age population and the total population (Figure 2), which indicates that our analysis is robust to the age range that we use.

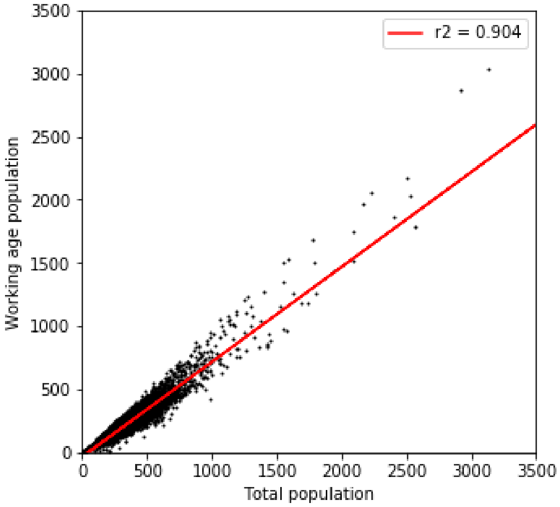

Figure 2: Working-age population compared to total population for all London Output Areas.

Routing data

In order to estimate the duration of trips for trips between stations, we require access to the Transit for London schedules of the services for each line. Although these schedules are publicly available, extracting and processing them was not possible given our time constraints. We therefore opted to use TripGo, a routing API which estimates travel time between any two locations [19]. Upon providing a set of parameters (origin coordinates, departure coordinates, departure time, and mode of transport), TripGo calculates the fastest paths to arrive from origin to destination.

Since we are using an API, we are limited in the number of server requests we can perform, which means that we must limit the number of routes we analyze. For this reason, in Section 4.4 we reduced our analysis to key hotspots. In the future, our method can easily be expanded to include all routes in the network.

3. Methodology

3.1 Station accessibility

A large component of our work involves finding the correlation between the station ridership and the daytime and nighttime population at a given station. However, in order to do so, we must model where transit riders are coming from. For this, we designate "station zones". All entrances to a station are assumed to be coming from passengers located within the station's zone.

In order to design the station zone, we make two basic assumptions. First, we assume that transit riders will enter the station that is closest to them. Second, we assume that riders walk a maximum of 500m to arrive at a station. With these two generalizations, we are able to create a zone for each station by intersecting the Voronoi tessellation of the network with a 500m buffer around each station (Figure 3).

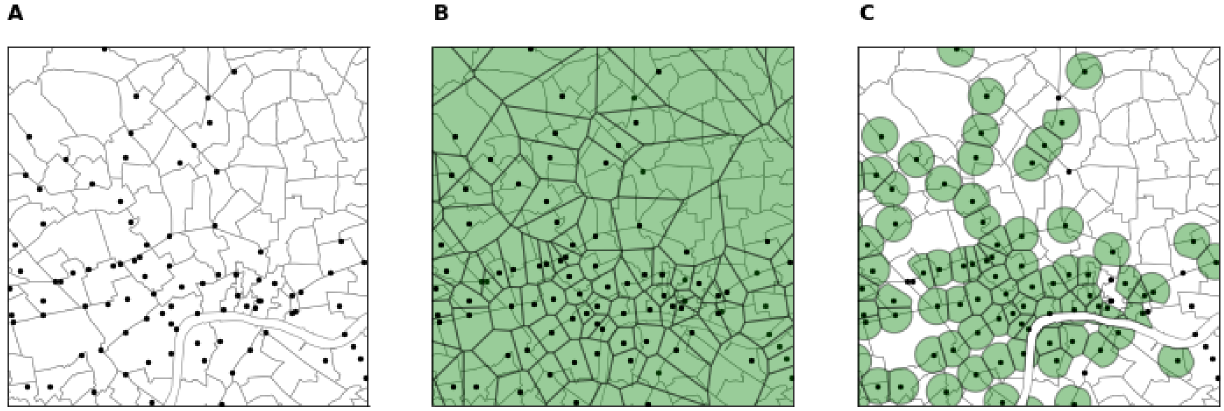

Figure 3: Creation of station zones. A portion of London is shown here. A) Locations of stations. B) Voronoi cells around each station. C) Intersection of Voronoi cells and 500m radius buffer.

Of course, our model for station accessibility is simplified. First, not every person entering a station is necessarily located within that station's zone. In reality, passengers may connect with a station via other modes of transport such as private vehicles, bicycles, buses, and overground rail. Under our accessibility model, passengers coming from completely different areas would still be considered to be living or working near a station simply because they connect with that station during their daily commute. We will show later how we eventually account for this "error" in our analysis, and use it to extract information about the transit network.

Another issue with our accessibility model is that it does not consider the transit lines to which the stations belong. The assumption that passengers will enter the station closest to them is reasonable when both stations are part of the same transit line, but when each station belongs to different lines, or to multiple lines, this assumption is no longer valid. For example, a passenger may be located 200m from one station on Line A and 400m from another station on Line B. Both stations are walking distance and depending on convenience, a passenger may enter either station. However, under our accessibility model, we assume the passenger will enter the station located on Line A because it is closer.

In addition to the weaknesses mentioned above, there are certainly many more ways in which we can improve the accuracy of our accessibility model, such as by using the true street distance rather than Euclidean distance for our buffer. However, we emphasize that our aim is not to maximize the accuracy of our model; there is already an active field of transit research which attempts to do so. We simply borrow key aspects from existing research in order to create a simplified model with minimal parameters which we can subsequently use when establishing the relationship between ridership and population distribution.

3.2 Extracting commuters

The time series of entrances into the station represents an aggregation of many different mobility patterns, not just commuting behaviour. For example, it may include passengers going to university classes or shopping. Therefore, in order to focus only on the commuting behaviour of riders, we require a mechanism to separate the commuting behaviour from other types of mobility, such as people going shopping or going back home after a night out. For this, we create a model which is based on typical commuting behaviour.

First, since the vast majority of commuting takes place on weekdays, we ignore the ridership patterns on weekends, corresponding to Saturday and Sunday. Second, since most work takes place during the day rather than during the night, we ignore all nighttime ridership, which we define as all station entrances occurring between 8PM and 6AM. After excluding this data, we are left with the time series of entrances into a station between 6AM to 8PM on a typical weekday.

Our next step is to differentiate the entrances representing commuting trips from other types of trips. For this, we separate the daytime trips into three categories:

1. *Workbound*. Passengers leaving home to go to work.
2. *Homebound*. Passengers leaving work to go back home.
3. *Miscellaneous*. All other mobility, such as shopping or going to appointments.

We model workbound trips as taking place in the morning (6AM to 11AM), and homebound trips as taking place in the afternoon (3PM to 8PM). For miscellaneous trips, we assume they take place mostly in between these times (11AM to 3PM). In order to reflect the fact that not all passengers in the morning and afternoon are daily commuters, we model the miscellaneous trips as linearly increasing and decreasing in the morning and afternoon, respectively (Figure 4).

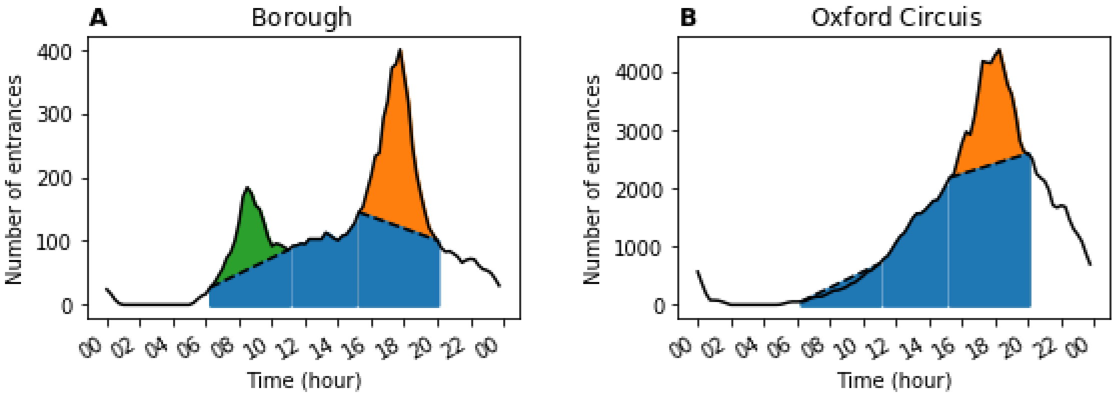

Figure 4: Classifying ridership. Riders during the day are classified into three categories: homebound (orange), workbound (green), and miscellaneous (blue). A) Typical result of classification; Borough Station is shown here. B) Example of a classification resulting in a negative workbound ridership; Oxford Station is shown here.

Under this model, the number of workbound riders can be found by subtracting the number of miscellaneous riders in the morning from the total number of morning riders:

$$R_{workbound} = \sum_{i=6AM}^{10AM} (R_i - R_{i_{miscellaneous}}), \qquad (1)$$

where $R_{\text{workbound}}$ is the number of riders going to work, R_i is the number of riders during a given time slot, and $R_{\text{miscellaneous}}$ is the number of riders which are expected to have miscellaneous motives (i.e. not homebound or workbound).

It should be noted that since the number of miscellaneous riders is a modeled term which is independent of the total number of riders, it is possible that subtracting the miscellaneous riders from the morning or afternoon riders results in the estimated workbound or homebound ridership being negative for some stations. Though we do not notice this for workbound ridership, we do notice this for homebound ridership, where several stations located in core business and areas experience such a large work-like characteristic that the off-peak traffic is higher than the morning traffic (Figure 4B). Although a negative ridership carries no physical meaning, the relative workbound ridership between different stations still holds. That is, stations with lower homebound or workbound values can be said to have a smaller amount of commuters compared to those with higher values.

4. Results

4.1 Characteristics of commuting transit

Before proceeding with our analysis of the transit data, we first show the significance of the commuting travel pattern. We specifically aim to show the symmetry of daily commutes and their dominance of the overall mobility taking place on the transit network.

Symmetry in our case means that trips in one direction are accompanied by a similar number of trips in the opposite direction. In order to examine this, we use the RODS dataset, which contains origin-destination data between all stations. We can examine the symmetry by calculating the correlation coefficient between opposite direction trips during different time periods (Figure 5). As expected, we see that the strongest symmetry occurs between morning and afternoon trips, followed by a symmetry between morning and evening trips. This suggests that commuting trips are highly symmetric: people who go to work in the morning take the same way to get back home in the afternoon.

This symmetry means that conclusions that we make about workbound ridership can also be applied to homebound ridership. We take advantage of this fact when performing subsequent analysis. In addition to its use in analysis, the symmetry of the commuting trips carries significance for transit planning, as it essentially doubles the effect of any shortcomings in the transit network, while at the same time offering double the returns for any steps taken to improve the service.

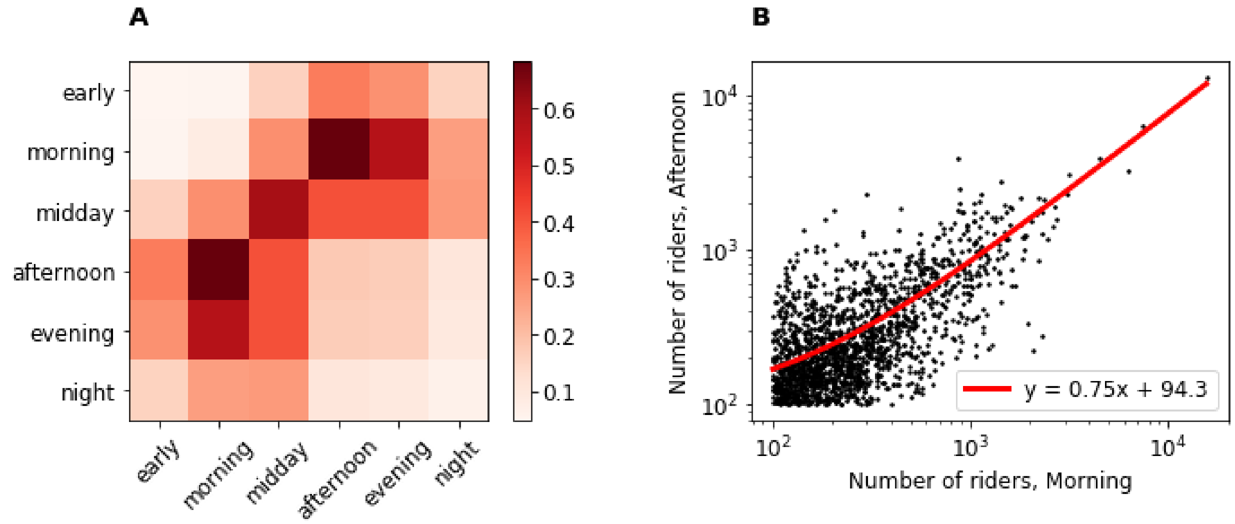

Figure 5: Symmetry of morning and afternoon riders trips. A) Correlation between trips in opposite directions for different times of day. For example, the bottom left square represents the correlation between trips made in one direction (e.g. from station i to station j) at night and in the opposite direction (from station j to station i) in the early morning. B) Number of trips between pairs of stations in the morning and in the opposite direction in the afternoon, for routes with more than 100 riders in each direction. Each point represents a pair of stations.

It is notable that the morning and afternoon routes are not just symmetric, but also represent the dominant form of mobility taking place on the transit network, in terms of sheer numbers. In the RODS dataset, symmetric morning-afternoon trips, which most likely represent daily commutes, represent 47% of all trips taking place on a weekday. Combined with its symmetrical properties, this exposes commuting mobility as a highly regular, everyday mobility which dominates weekday urban activity.

4.2 Transit and urban structure

In this section we aim to show the relationship between the ridership pattern and the urban structure. Though attempts have been made to do so using origin-destination trajectories, we show that even with basic information such as the number of entrances over the span of a day, we are able to extract significant information about the work-home dynamics of the city.

One would expect stations in residential areas to experience a large proportion of entrances to stations in the morning, while stations in business districts to experience a large proportion of entrances in the afternoon. In order to characterize this quantitatively we introduce the concept of a *residentiality index*, which we define as the difference between the peak number of entrances in the morning the number of entrances in the afternoon. This is normalized by the number of riders in the off-peak in order to remove the effects of absolute traffic volume and allow for comparison between stations:

$$residentiality = \frac{\max(R_{morning}) - \max(R_{afternoon})}{mean(R_{offpeak})}, \qquad (2)$$

where $R_{morning}$, $R_{afternoon}$, and $R_{offpeak}$ are respectively the number of riders in the morning, afternoon and offpeak.

This index can be interpreted as a measure of how residential the ridership behaviour of the station is. Stations with large positive residentiality represent stations with large morning peaks and small afternoon peaks, indicating a large proportion of workbound riders, while stations with a large negative residentiality indicate a large proportion of homebound riders. Stations with ratios close to zero show roughly equal amounts of workbound and homebound riders. As this index is normalized, it is independent of the actual volume of riders.

With just this index, we are able to generate important information about the overall urban environment of London. For example, by ranking the stations according to residentiality, one can see that there are many more residential stations than there are work-like stations (Figure 6A). Moreover, if we remove the normalization of the index and therefore also consider the absolute volume of each station, we see that not only are there many more residential stations but also that these stations have less traffic than the business stations (Figure 6B). This information is consistent with the “many-to-one” characterization of a typical city [5], where many residential areas feed a small number of central business districts. Though beyond the scope of our work, one could imagine a simple metric, such as the ratio of positive-residentiality stations to negative-residentiality stations, which could provide an abstract representation of the centrality which can be used for cities worldwide.

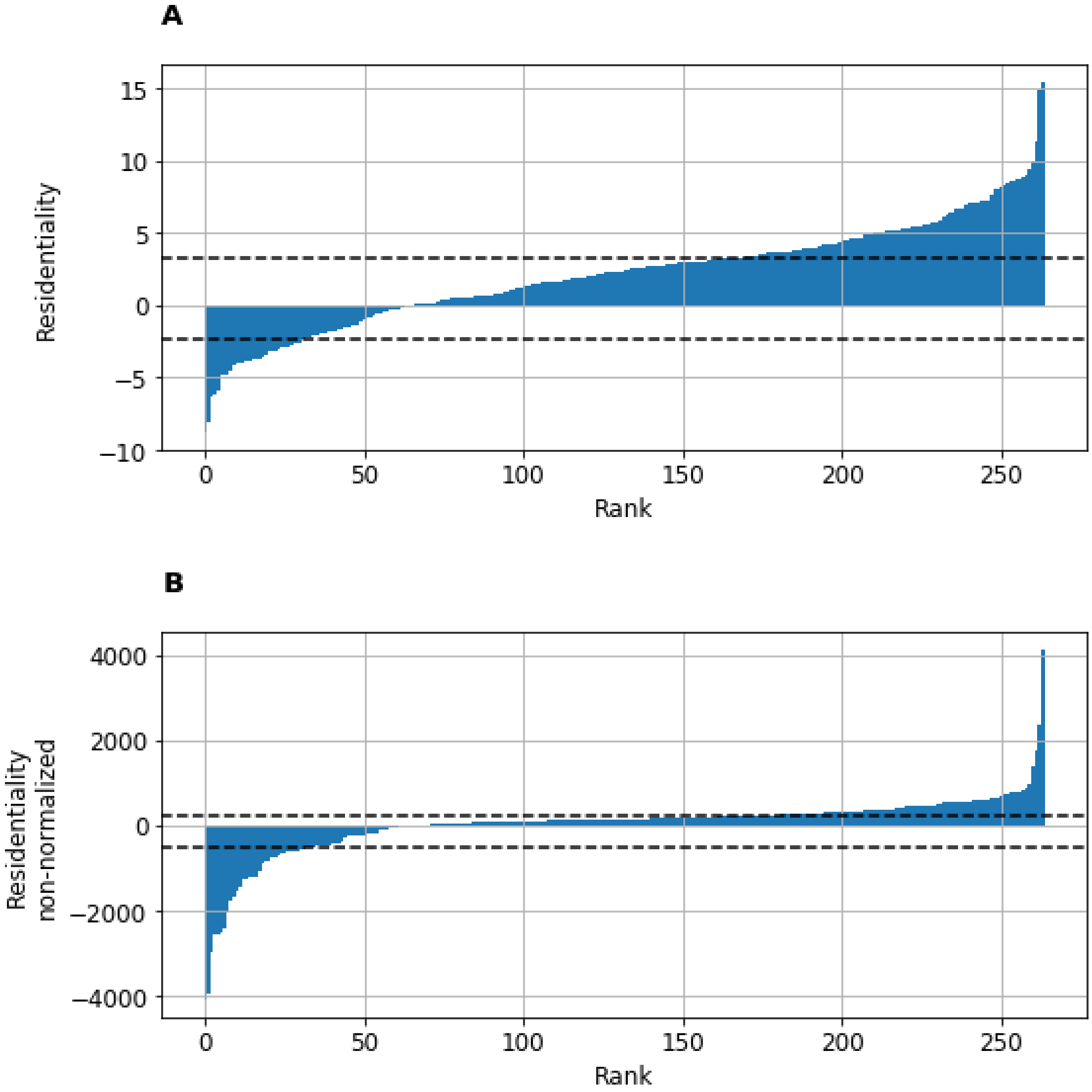

Figure 6. Ranking of stations by residentiality. Dotted lines show the median level. A) Stations ranked by normalized residentiality index. B) Stations ranked by non-normalized index, which includes information on absolute traffic volume.

In addition to the absolute number of residential and non-residential stations, there is a spatial distribution to the residentiality index. We show this by introducing a simple threshold for the residentiality: stations above the threshold are grouped separately from those below the threshold, with the two groups representing residential and business stations, respectively. When measuring the average distance between the stations, we find that the business stations are much closer together than the residential stations (Figure 7). This indicates that the residential areas are spread out on the outskirts of the city, while the business districts are clustered together. In the case of London, it reveals the dense urban core that is the City of London, surrounded by a residential periphery.

In addition to revealing the central business district (CBD) of London, the residentiality index can reveal other commercial areas in the city. For example, our method exposes a strong work-like ridership pattern in Canary Wharf Station, corresponding to the large financial district in the Isle of Dogs area. Since our metric is normalized for traffic volume, it can also reveal areas which, though relatively low in ridership, are heavily business-oriented. For example, our metric indicates White City Station as being commercial, which corresponds to the location of various British Broadcasting Corporation (BBC) buildings and to the location of Westfield Mall, the largest mall in Europe [20]. Thus, our metric is able to distinguish the polycentric nature of London, where the large CBD is surrounded by smaller commercial areas [5].

By creating a threshold for the residentiality index, we are essentially introducing a classification problem. By varying the threshold, one is varying the classification between residential stations and business stations. The natural extension of this is therefore to find a critical threshold which best reflects the actual work and home locations in the city. Though further research can be done to find the ideal threshold, we propose a possible critical threshold as the value of the residentiality index where the difference between the inter-station distances of home and business stations is greatest (i.e. where the two lines in Fig. 7A are farthest apart from each other). This is not only an intuitive threshold to understand and use, but also is nonparametric and therefore could be easily applied to other cities and transit networks.

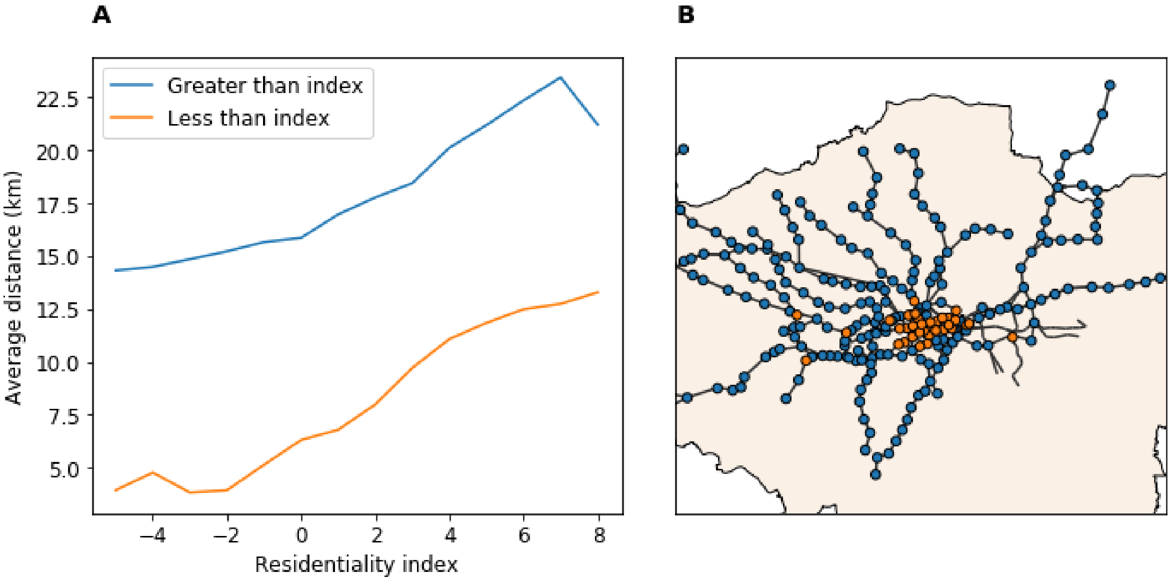

Figure 7: Distances between stations. A) The distances between stations above and below a residentiality threshold. B) A visualization of the distribution of work and home stations at the threshold where the difference between the two inter-station distances is greatest.

The residential index we develop does not require any origin-destination trajectory information; only data about the daytime entrances into each station over the course of a day is necessary to determine the residential nature of a station (and its surrounding area). Therefore, with only this data, our results show that we are able to describe the urban structure and exposes residential and business centers.

4.3 Accessibility analysis

Considering the relationship between the ridership pattern and the urban environment, one would expect a similar relationship between ridership and the population distribution. For example, one should see a strong correlation between the number of workbound riders in a station and the population within the station's zone. This is quite intuitive: the number of people taking the train in the morning from a given station should be proportional to the number of people living within the proximity of that station. However, we see that although the two are positively correlated, the relationship is quite weak and prone to outliers (Figure 8).

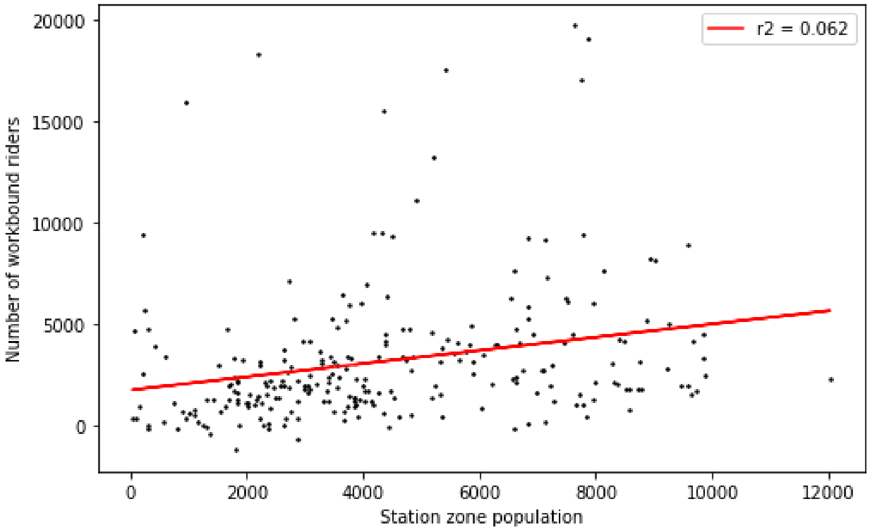

Figure 8. Population of station zones compared to workbound ridership. Though positively correlated, the regression is error-prone and contains several significant outliers.

The error in this ridership model can be explained by examining its mathematical form more closely. By regressing workbound ridership against population, we are essentially modelling ridership as proportional to the population:

$$R_{workbound} = \alpha P, \qquad (3)$$

where $R_{workbound}$ is the workbound ridership, P is the residential population within the station zone, and α is the proportionality constant.

This model reflects some assumptions we make. Namely, we assume that passengers enter the station that they live nearest to, and we assume that, of those passengers who do enter the station nearest to them, that α, the proportion of passengers taking the train compared to other modes of transport, is constant across all stations. In reality, passengers entering a station may come from far away, having walked or used other modes of transport to get to their desired station. Moreover, the proportion of passengers using the transit system is known to depend heavily on socioeconomic factors, and therefore is not constant across all stations. These differences from our assumptions are responsible for the over- and underestimations of our linear regression in Fig. 8.

With the data we have, we cannot compensate for these errors. However, we can define a new, theoretical model which more realistically reflects the ridership at a station:

$$R_{workbound} = \alpha_s P + R_{fromElsewhere} - R_{toElsewhere}, \qquad (4)$$

where α_s is the station-specific proportion of the residential population using the transit network, $R_{fromElsewhere}$ is the number of workbound riders coming from areas outside the zone in order to enter the station, and $R_{toElsewhere}$ is the number of workbound riders leaving their zone to go to another station. It follows that the difference between these two models represents the error in our earlier regression:

$$\epsilon = R_{fromElsewhere} - R_{toElsewhere} + (\alpha_s - \alpha_{mean})P. \qquad (5)$$

Thus, the error actually represents some critical properties of the transit system. A large positive error can represent one of two things: a large number of riders coming from other areas, or a larger than expected percentage of inhabitants taking the train. On the other hand, a large negative error represents the opposite: many inhabitants taking the train somewhere else, or fewer than expected inhabitants taking the train at all. Essentially, the error in the regression is not an error at all, but rather a description of how well the station performs. Stations which perform well attract passengers from other areas, whereas stations which perform poorly lose their own passengers, perhaps due to bad service or inconvenient connections.

Given how informative it is, we propose the regression between population and workbound ridership as a method of examining transit accessibility. Without the use of any socioeconomic information or an analysis of the transit service, our method is able to highlight locations which suffer from poor ridership, which in turn are indicative of unsatisfactory transit service. Our method therefore serves as a transit planning tool that is both intuitive and easy to apply with limited data.

It is notable that with this method, strong work hubs such as Oxford Station would be considered to be poorly performing due to their low morning ridership (Figure 4B). This would be misguided, since we know that these stations are well connected and used widely in the afternoon. Though we cannot say with certainty why the ridership is low in the morning, we suspect that this may be due to the fact that work locations are so close to these stations so as not to necessitate travel via transit. Regardless of the reason, for a complete transit accessibility analysis, one would need to separate the poorly performing stations from "false positives" such as Oxford Station. Though beyond the scope of our work, we believe that this can be done by accounting for the afternoon ridership in addition to the morning ridership, or by compensating for the number of work locations in a city. Both of these methods would filter out work hubs, which are located in central areas where many people work and use the stations to get home in the afternoon.

4.4 Hotspot analysis

Complementary to investigating the underused stations of the transit network, in this section we examine the most popular stations and routes. We show how the diurnal ridership pattern can be used to identify the residential and business hotspots of the transit network. We then proceed to show that, after identifying these hotspots, we can expose deficiencies in the service between these hotspots.

We begin by identifying the residential and work hotspots in the network. We define these hotspots as being the stations which are major sources or destinations commuting trips taking place on the network. Given this definition, it follows that we can find these hotspots by identifying the stations which have high amounts of homebound or workbound riders. Residential hotspots correspond to those stations which have a high number of workbound riders, while work hotspots correspond to stations with a high number of homebound trips. This is different from the residentiality index we developed earlier in Section 4.2. First, the residentiality index is normalized for the volume of traffic at the station, while the hotspot metric includes traffic volume in order to separate the high-traffic stations from the low-traffic stations. Second, whereas the residentiality index necessarily creates a trade-off between residentiality and business-like ridership, the hotspot measure does not, so as to reflect the fact that a station can simultaneously be a major source and destination of daily commutes. For example, a major rail hub such as Waterloo station is a common transit connection both in the morning and in the afternoon.

Once the volume of workbound and homebound traffic is determined, the high-traffic hotspots must be separated from the rest of the stations. The detection of hotspots and urban centers is a common research problem in urban studies, and often it relies on defining a threshold separating urban centers from other areas [7][21]. We use the simplest threshold, which is to take the top 20 residential stations and the work stations (Figure 9). Though choosing the top 20 stations is rather arbitrary, we note that the actual method of hotspot identification is not important in our case, and we only use it in order to reduce the number of requests on the routing API, which has a limited quota. In the future, the analysis would be performed on the entire network, and therefore no hotspot identification is necessary.

When we examine the distribution of residential and work hotspots, we can see it takes on a monocentric structure, reflecting the dense business district of central London and the residential areas surrounding it. We also note that the hotspots tend to be stations with National Rail connections, such as Waterloo, Brixton, and Stratford stations. Though the stations themselves may not have many residents in the vicinity, they are important connections which collect residents coming from far away via train. This highlights the fact that these station hotspots do not necessarily coincide with actual work and home urban areas, but rather with high-traffic network nodes.

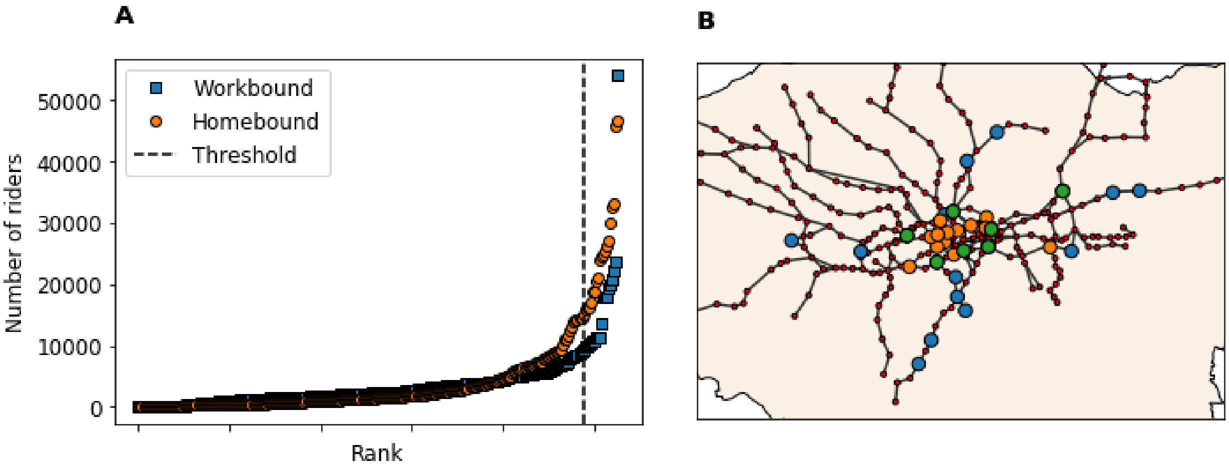

Figure 9. Hotspot extraction via thresholding. a) Ranking the stations by number of workbound and homebound riders. b) Visualization of the home (blue) and work (orange) hotspots. Stations which are both home and work hotspots are shown in green.

With the hotspots identified, we can now examine how well-suited the transit network is to the commuting flow between hotspots. Specifically, we show that by examining the commuting flow between hotspots, we are able to find the most underperforming commuting routes. For this, we introduce a new metric, which we call *potential for time savings* (PTS), which represents the amount of time that can be saved cumulatively for all passengers on a given route, for a given increase in speed:

$$PTS_{ij} = \frac{R_{ij}}{v_{ij}}, \qquad (6)$$

where R_{ij} is the number of riders on a given route from station i to station j, and v_{ij} is the straight-line speed of the service from station i to station j.

The PTS does not literally represent the amount of time that can be saved. Instead, it represents the relative ease or difficulty with which time can be saved. If a service between two hotspots is very popular but is already very fast, that route is probably already optimized and would therefore have a lower PTS than a slow route. The reason we choose to use the speed of the route instead of the absolute time it takes to travel from i to j is that the speed is not dependent on the distance of the route. For example, a reduction travel time of 2 minutes for a given route represents a smaller proportion of the total travel time for a long-distance route than it does for a short-distance one, and therefore is most likely much easier to accomplish on a longer route. By focusing on speed, we are able to account for time delays such as transfers and poor connections, while being able to represent the potential improvements that can be made to a route regardless of distance. Though speed is not a perfect measure for transit service (for example, it ignores the headway, or frequency, of the scheduled trips), it serves as an approximation for the quality of service.

In order to determine the PTS of each commuting route, we must determine the ridership and speed of each route. We begin by estimating the ridership. For this we use the RODS dataset, which, unlike the Passenger Count dataset, contains origin-destination information. Using this dataset, we create an origin-destination (O-D) matrix representing the volume of traffic, in number of riders, between all residential and work hotspots (Figure 10). Essentially, this conveys the demand for a given route, and serves as a measure for how many people would be affected by an improvement (or impairment) to the service of the route.

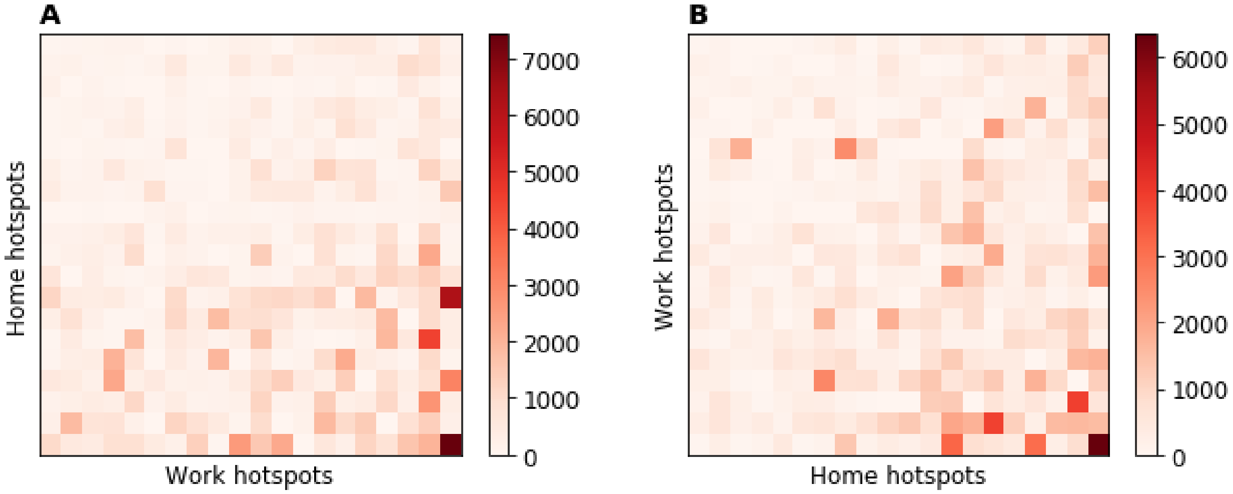

Figure 10: Number of riders traveling between hotspots. a) Riders traveling during morning peak. b) Riders traveling during afternoon peak.

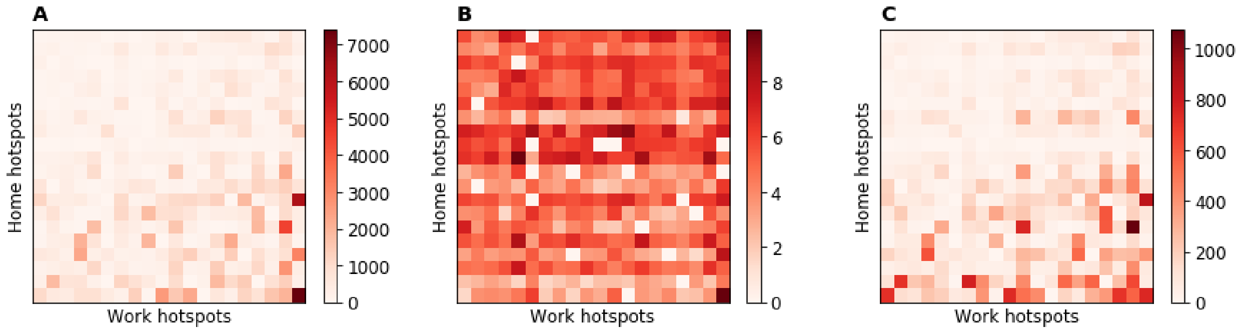

Figure 11: Determining PTS for morning transit. A) Number of riders between hotspots in the morning. B) Average speed (in m/s) for routes between stations, based on Euclidean distance between the stations. C) Potential for time savings, determined by dividing the first two matrices.

Having determined the traffic volume (Figure 11A), we proceed to calculate the speed of each route. For this we use the TripGo API, which provides the amount of time required to travel from i to j. As there are numerous scheduled trips for each route, we take the minimum time of these trips, in order to represent the maximum speed of the route which is currently being achieved. We divide the straight-line distance from origin to destination by the time in order to get the speed. As with ridership, we represent the speed of each route in an O-D matrix (Figure 11B).

Once both the ridership and the speed are determined, we proceed to estimate the potential for time savings of each route, which is simply a matter of dividing the two terms (Figure 11C). From the resulting matrix, we can see several routes which, due to a combination of heavy traffic and suboptimal speeds, result in a large amount of extra time spent cumulatively by the passengers. These routes could be prime candidates for service improvement in the future, as they represent the greatest returns in terms of time reduction for a given improvement in speed. One could imagine, for example, improving the speed of a specific route by creating an express line, or by reducing boarding and alighting times.

5. Conclusions

In this work we showed that the number of entrances into transit stations over time can be used to extract information about the urban environment and the transit performance. Specifically, we introduce several novel analytical methods which extract metrics from the ridership pattern which require few parameters and which can be expanded to transit networks worldwide. First, we show the inherent connection between the daily urban commutes and the number of riders in the morning and afternoon. Based on this relationship, we introduce a *residentiality index*, which characterizes how residential or work-like a station is based on the number of riders throughout the day. We then proceed to show that the residentiality index is able to reflect the urban structure, identifying key business districts throughout the London region. Essentially, we show that entry-only ridership data can be used to identify urban home and work dynamics. We introduce a possible non-parametric thresholding of the residentiality index based on the inter-station distances of home and work stations.

Second, we use this relationship between urban structure and transit ridership in order to analyze the accessibility of the transit network. We introduce a nonparametric method which is able to extract transit gaps based on the discrepancy between population and the number of workbound riders. Instead of estimating demand and supply using multiple parameters and many datasets, we can determine station accessibility using just two datasets: the census population and the number of entrances into the stations. These two datasets are easily accessible to transit agencies and therefore our method can be easily adapted to other transit networks worldwide.

In our third and final analysis, we used ridership information in order to estimate the performance of the network routes. We use our model to extract hotspot stations, which represent the stations with the greatest number of workbound or homebound riders. We then introduce a metric, *Potential for Time Savings* (PTS), which combines ridership data with timetable information in order to represent how much time can be saved for each trips between hotspots stations. We show that with this metric, we are able to extract poorly-performing routes, which would normally go undetected.

Our analysis serves as a preliminary study highlighting the utility of entry-only ridership data. We show that even without origin-destination trajectories, which have become commonplace in urban studies, important information about urban structure can be extracted. As such, the methods we introduce can be expanded to other transit networks much more easily than existing methods, and can therefore become a critical tool for future urban and transit planners.

5.1 Future Work

Our analysis depends on being able to accurately extract the number of homebound and workbound riders from only the pattern of entrances over the course of a day. For this we defined a model, which we presented in Section 3.2. However, we have not done a rigorous analysis to validate our model. One possible way to do so in the future is to compare the modeled number of workbound and homebound riders to the number of symmetrical trips made in the morning and afternoon. That is, if our model is valid, the modeled number of commuters should be similar to the number of passengers who make a trip in the morning and an opposite-direction trip in the afternoon. With this validation, we expect that the morning and afternoon hours used may need to be adjusted. For example, by customizing the morning and afternoon hours for each individual station, we can compensate for the fact that stations far away from the center will have commutes that start earlier than stations closer to the center, which are naturally closer to the work locations.

Another limitation of our work is the intermodal transport that takes place in the city. In our analyses, we assume that passengers enter the station closer to them when in reality, passengers may walk large distances or take other means of transport in order to get to their desired station. This results in large outliers when relating the ridership to the local population (Figure 8), where some stations, such as Waterloo Station, may receive passengers from far away locations. By incorporating National Rail and bus ridership data in the future, we would be able to decouple these non-local effects. For example, we would be able to detect passengers taking the regional rail transit before they connect with an underground station. This would increase the correlation between ridership and local population, therefore improving the correspondence between the conclusions we make about the network and the actual urban environment. For example, any estimate we currently make of the residentiality of Waterloo Station only reflects the residentiality of the transit node in the network, not of the actual urban area, because the ridership represents riders from far away areas.

When analyzing urban structure, we defined a residentiality index as a metric that can be used to describe cities. We also introduced a possible method of classifying urban areas as work or home based on a threshold of their residentiality index. However, before this can be applied throughout cities worldwide, further work must be done in finding the threshold which best represents the true separation between home and work. For example, by incorporating information about business locations, one would be able to validate the true balance between home and work throughout the city, and use it to determine a realistic residentiality threshold. Once a the threshold is described, there would be a consistent definition for defining urban areas as home or work. From this, research can be done to establish the usefulness of the residentiality index in describing and classifying cities. For example, one may find that there is a global pattern in the balance between the number of home and work locations, or that the location of home to work hotspots can be used to categorize cities.

A natural extension of our work is to apply the methods we introduced to other transit networks. Currently, our work is restricted to the London Underground network. Extending our work to other cities would be a major step toward validating our models and results. Given the wide availability of transit station entrance data, we expect our work to be easily adapted to other transit datasets. Ultimately, we expect the methods we introduced in this work to become a standard tool for urban planners and transit agencies.

References

[1] Park, J. Y., Kim, D.-J., & Lim, Y. (2008). Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea. *Transportation Research Record: Journal of the Transportation Research Board*, 2063(1), 3–9. https://doi.org/10.3141/2063-01

[2] Taylor, B. D., Miller, D., Iseki, H., & Fink, C. (2009). Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. *Transportation Research Part A: Policy and Practice*, 43(1), 60–77. https://doi.org/10.1016/j.tra.2008.06.007

[3] Gutiérrez, J., Cardozo, O. D., & García-Palomares, J. C. (2011). Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19(6), 1081–1092. https://doi.org/10.1016/j.jtrangeo.2011.05.004

[4] Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, 53, 19–35. https://doi.org/10.1016/j.compenvurbsys.2015.02.005

[5] Roth, C., Kang, S. M., Batty, M., & Barthélemy, M. (2011). Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE*, 6(1), e15923. https://doi.org/10.1371/journal.pone.0015923

[6] Zhong, C., Arisona, S.M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28:11, 2178-2199, DOI: 10.1080/13658816.2014.914521

[7] Louail, T., Lenormand, M., Cantu Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., … Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. *Nature Scientific Reports*, 4(1). https://doi.org/10.1038/srep05276

[8] Noulas, A., Mascolo, C., & Frias-Martinez, E. (2013). Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments. *IEEE 14th International Conference on Mobile Data Management*, Milan, 2013, pp. 167-176. doi: 10.1109/MDM.2013.27

[9] Barry, J. J., Newhouser, R., Rahbee, A., & Sayeda, S. (2002). Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board*, 1817(1), 183–187. https://doi.org/10.3141/1817-24

[10] Zhao, J., Rahbee, A., & Wilson, N. H. M. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided*

Civil and Infrastructure Engineering, 22(5), 376–387. https://doi.org/10.1111/j.1467-8667.2007.00494.x

[11] Foth, N., Manaugh, K., & El-Geneidy, A. M. (2013). Towards equitable transit: examining transit accessibility and social need in Toronto, Canada, 1996–2006. Journal of Transport Geography, 29, 1–10. https://doi.org/10.1016/j.jtrangeo.2012.12.008

[12] Stanley, J., & Lucas, K. (2008). Social exclusion: What can public transport offer? Research in Transportation Economics, 22(1), 36–40. https://doi.org/10.1016/j.retrec.2008.05.009

[13] Fransen, K., Neutens, T., Farber, S., De Maeyer, P., Deruyter, G., & Witlox, F. (2015). Identifying public transport gaps using time-dependent accessibility levels. *Journal of Transport Geography*, 48, 176–187. https://doi.org/10.1016/j.jtrangeo.2015.09.008

[14] Bocarejo S. J. P., & Oviedo H. D. R. (2012). Transport accessibility and social inequities: a tool for identification of mobility needs and evaluation of transport investments. *Journal of Transport Geography*, 24, 142–154. https://doi.org/10.1016/j.jtrangeo.2011.12.004

[15] Currie, G. (2010). Quantifying spatial gaps in public transport supply based on social needs. *Journal of Transport Geography*, 18(1), 31–41. https://doi.org/10.1016/j.jtrangeo.2008.12.002

[16] Transport for London. (2017). London Underground passenger counts data. Retrieved January 10, 2019, from https://api-portal.tfl.gov.uk/docs

[17] Transport for London. (2017). Rolling Origin & Destination Survey (RODS). Retrieved January 10, 2019, from https://api-portal.tfl.gov.uk/docs

[18] Office of National Statistics. Census Output Area population estimates – London, England. (2018). Retrieved January 10, 2019, from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/censusoutputareaestimatesinthelondonregionofengland

[19] SkedGo Pty Ltd. TripGo API. Retrieved February 1, 2019, from https://developer.tripgo.com/

[20] Hipwell, D. (2018). Westfield is now Europe's biggest mall. *The Sunday Times*. Retrieved from https://www.thetimes.co.uk/article/westfield-is-now-europes-biggest-mall-zzmmkn6fn

[21] Giuliano, G., & Small, K. A. (1991). Subcenters in the Los Angeles region. *Regional Science and Urban Economics*, 21(2), 163–182. https://doi.org/10.1016/0166-0462(91)90032-i