Finding Pearls in London's Oysters

JON READES[1], CHEN ZHONG[1], ED MANLEY[2],
RICHARD MILTON[2], & MICHAEL BATTY[2]

Public transport is perhaps the most significant component of the contemporary smart city currently being automated using sensor technologies that generate data about human behaviour. This is largely due to the fact that the travel associated with such transport is highly ordered. Travellers move collectively in closed vehicles between fixed stops and their entry into and from the system is unambiguous and easy to automate using smart cards. Flows can thus be easily calculated at specific station locations and bus stops and within fine temporal intervals. Here we outline work we have been doing using a remarkable big data set for public transport in Greater London generated from the smart card data called the ***Oyster Card*** *which has been in use for over 13 years. We explore the generic properties of the Tube and Overground rail system focussing first on the scale and distribution of the flow volumes at stations, then engaging in an analysis of temporal flows that can be decomposed into various patterns using principal components analysis (PCA which smooth out normal fluctuations and leave a residual in which significant deviations can be tracked and explained. We then explore the heterogeneity in the data set with respect to how travel behaviour varies over different time intervals and suggest how we can use these ideas to detect and manage disruptions on the system..*

[1]Department of Geography, Kings College London, King's College London, Strand Campus,
London WC2R 2LS, UK; j.reades@kcl.ac.uk, c.zhong@kcl.ac.uk
[1]Centre for Advanced Spatial Analysis, University College London, 90 Tottenham Court Road,
London W1T 4TJ, UK; r.milton@ucl.ac.uk, e.manley@ucl.ac.uk, m.batty@ucl.ac.uk

BIG DATA, AUTOMATION AND SMART TRANSIT

Automation in transit systems is the most visible sign of how the city is being transformed to enhance the travel experience and efficiency of movement (Batty et al., 2012). There are many ways of achieving this but one of the most significant is the use of smart cards for 'fully automatic fare collection'. These smart cards usually contain the value that the consumer has agreed to load onto the card, they meet stringent requirements for anonymity and security, and their use is such that by tapping in and out of an automated system, correct payments are ensured. Smart cards like this, in fact, go back to the late 1960s and rapid progress in their development was achieved in the 1970s and 1980s when they first made their appearance as phone cards in France. Different varieties of credit card were then emerging, and by 1984 in places like Hong Kong, stored value cards for use on their new Mass Transit Railway (MTR) were introduced. By the mid-1990s, contactless cards came onto the scene, first in Seoul with the UPass card, and then in Hong Kong where they introduced the Octopus card which was then extended to other purchases in the local retail system.

Several other cities followed, but one of the most comprehensive rollouts was in London where, in 2003, the first cards were introduced on the underground ('Tube') system. These are called 'Oyster' cards – partly in tribute, it would seem, Hong Kong's Octopus card – for the official reason that the Oyster card protects its 'pearl' – the stored value – in a 'hard shell'; hence, the name which we have used in the title to this paper. Our particular interest in these 'pearls' is not in their value but in the raw data that can be extracted which covers 'where' the owner of the card taps in and out, the times 'when' this takes place, and the status of the card used. Carrying more than 1.3 billion passengers per year, the London network generates a vast quantity of data – 'big data' in the terminology used here – and our intention in this paper is to explore and critique the generic properties of transit in London using a subsample of this 'big' data.

The use of automated fare collection systems to retrieve travel data has a relatively long history, and there are now quite good reviews of applications from Montreal (see Morency, et al., 2006, and Pelletier et al., 2011) and Chile (Munizaga, et al., 2014). A series of reviews are also currently available on the web (such as Alguero, 2015). At the margins, much of this work overlaps with, and merges into, network- and location-led research on social media and communications from smart phones (see Gonzalez et al., 2008 and Ahas et al., 2015), as well as from applications such as Twitter and Foursquare (see Noulas et al., 2012). However, although there are important constraints – which we will discuss in more detail shortly – on the accuracy of transit data, smart card systems nonetheless hold out the promise of a kind of 'gold standard' in activity, event, and flow research since the system is much more 'contained', and more broadly representative in terms who it 'captures', than many telecoms networks.

The willingness of Transport for London (TfL), the agency that manages the public transport system in London, to engage with researchers has resulted in carefully curated tranches of Oyster data being made available, subject to local licenses. Several academic groups have made use of this, notably the MIT Transport group led by Nigel Wilson who have produced a number of analyses of flows, services and reliability on the Tube (see for example Gordillo, 2006). Our own work began some 5 years ago when we used the data to cluster the main station hubs (Roth et al., 2012) and since then Silva et al. (2015) and Williams and Musolesi (2016) have produced additional analysis on statistical properties of the data which reveal the structure of travel demand and the relationships to shocks and disruptions on the system. We have also explored the degree of heterogeneity in traveller profiles at Tube stations (Manley et al., 2016) and variability in travel arrival times by comparing the London system to similar ones in Singapore and Beijing (Zhong et al., 2016).

Broadly, the research with smartcards falls into four areas of focus: systemic descriptions, inferred use and activity mining, mobility flows, and disruption modelling. Our contribution here is very much in the vein of the first category: an attempt to describe the generic properties of the transport system as revealed by the data. In many cases, that is a first step to understanding how such data can be synthesised to reveal activity and land use location patterns from a detailed analysis of movements and their integration with related data sets. The third and fourth areas are often also interlinked since the generation of trip patterns– origin and destination movements – from tap-in tap-out data is obviously integral to the area of predicting the impact of disruptions – from crowding and other extreme events – on the nodes and links of the network. Of less immediate interest, since many researchers are still struggling to make the most of these new data, is a fifth area dealing with the use of such data for control and management.

In terms of general properties of transit systems from such data, most of the work on different applications in places such as Shenzen (Gong et al. 2012), Beijing (Long and Thill, 2015), Santiago (Munizaga et al., 2012), Singapore (Zhong et al., 2015), and London (Gordon et al., 2013) deal with the general properties of these systems as well as ways of scaling trip movements to produce information about the locational activities associated with the places where people access the transit systems in question. In terms of extracting more detail on movement, there has been less progress because much of this depends in linking such data sets to cognate data and this requires a common key to enable integration of quite different data sets. In work on disruption, much of this is being pursued in the network science domain building on ideas about multiplexing, modal split and the resilience of the networks themselves. These are all features we will point the reader to in the analysis that follows.

In this paper we will first outline the salient characteristics and properties of the London rail transit system exploring how its stations capture movement. We will simplify the data set and handle only tap-ins to the system but this gives us a good sense of temporal and spatial changes from which we then derive profiles of traveller behaviour using the equivalent of statistics that measure the concentration of flows over the working day. We then conduct a series of analysis that enable us to extract components of variation (principal components) from the data and this gives us some sense of differences between time periods and station hubs over the week-long data set that we use in this example. We then examine briefly the variability of the data and imply some analysis of disruption to give the reader a glimpse of what is possible with data such as this. In fact, our analysis barely touches the iceberg of big data in this context for all we do is look at hubs not flows and at how the system generates changes in scale as travellers move differently through the working day and the weekend, in different spatial clusters within the system.

UNDERSTANDING THE OYSTER SYSTEM AND ITS DATA

It is tempting to think that bespoke digital smart card ticketing system like Oyster that neatly captures entries and exits – colloquially known as tap-ins and tap-outs – from the transport network, would allow new insights into travel behaviour at the city scale. However, this would be a fundamental misapprehension of the function of the Oyster system: it is not a travel analysis system at all, it is a ticketing system supporting a mix of pre- and post-travel charging mechanisms that seek to only ensure that users are charged the correct amount for their journey. For all such systems, analytical applications are, in most meaningful senses, an afterthought. To a very large extent, this is a feature of much of the big data that is explored in this special issue.

From this basic understanding flows a series of consequences: the data set tracking Oyster 'events' not only incorporates non-travel activities such as 'topping up' (adding stored value) and the application of penalty charges (after a time-out period), it also includes 'out-of-station interchanges' (when a user exits a station but is allowed a free transfer) and 're-entries' or 're-exits' (such as occur when a user enters a station, decides it is over-crowded, and exits in search of an alternative transport mode), and intermediate 'validation' taps (indicating that the user is bypassing a more expensive zone such as the central Zone 1 while journeying between stations that are also reachable via a Zone 1 route). Cumulatively, at the city scale even comparatively rare events – not many people exit a system only to re-enter seconds later – can nonetheless generate significant volumes in the overall set of events captured by TfL.

Compounding this challenge is the fact that stations themselves are surprisingly complex entities; Paddington Station, for instance, actually consists of three separate 'units' as identified by their National Locator Code (NLC) identifiers: the mainline station with services to Reading and beyond; the Circle and District part of the Underground; and the Hammersmith & City section (now also a 'Circle' line part) of the Underground. Then, of course, some stations carry both mainline and TfL-operated trains on the same platforms, leading to the question of who is allocated the revenue from that trip. Even intercity trains accept some Oyster-like cards such as the over-60's Freedom Pass which is operated by TfL. Furthermore, since the charging system was designed to work without a network connection, there is no guarantee that the timing information provided by the ticketing device is accurate and so events can easily appear to be out-of-synch, with tap-outs preceding the tap-in for an individual journey.

To the Oyster card, none of these issues is a problem: the system simply applies the charging rules defined by the operator in order to void, charge, cap, or refund a ticket. But to us – as researchers and transport analysts – this flow of events emphasises the extent to which post-hoc analysis of user activity is underpinned by a set of 'business rules' that are ultimately arbitrary and guided principally by decisions as to what is most appropriate given the analytical context. To provide a sense of how just how complex the system is, the Oyster data feed (on which this analysis is based) contains no fewer than 52 fields, including details on the memory 'slot' where the ticket is stored, fare adjustments, user-type, and whether or not the card was successfully updated by the event. Questions emerge such as 'Should the researcher take into account indirect location and travel time information provided via unsuccessful taps?' There is, simply, no one answer to such a question and this aspect of smart card analysis is often overlooked by 'big data' enthusiasts: attempts to apply lessons drawn from one charging and transport context – i.e. set of business rules and operational limitations – to an entirely different raises major challenges to the robustness of the analysis.

For this analysis, we have opted to work with three sets of derived activities to do with the rail based on the Overground, Docklands Light Rail, and Underground (Tube) networks. These 'modes' are of particular interest since they (usually) provide both tap-ins and tap-outs from which a geographical dimension to travel activity can be inferred. This gives us access to the spatial and temporal structure of the heavy, light, and underground rail networks as a whole: at the system and station levels which we deal with in the next section, we have focussed on valid tap-ins alone since they do not require us to match each entry with a valid exit and are commonly available in other public transit systems (e.g. New York, Paris, Singapore, Beijing etc.). 'Origin/Destination' data could be derived for activity in London, Beijing and Singapore, but probably not for New York or Paris because of decisions around charging, zones, and even the hardware deployed on each network.

For this analysis we chose a week when there was only routine disruption and no extraordinary events. This is the week starting Monday 2nd July 2002, which takes in all five weekdays and the two weekend days to Sunday 8th July. During this period, there are a total of 17.9 million tap-ins and 19m tap-outs from 296 stations, with an average weekday tap-in of 2.1m and tap-out of 2.2m, and weekend tap-ins of 0.47m and tap-out of 0.51m. An immediate problem is that the number of tap-outs exceeds the tap-ins by 1.2m (6%), which is a reflection of the fact that gates may be left open to ease congestion, and that commuters with weekly or monthly passes are not required to tap-in or out since their ticket is already 'valid' for the journey. In fact the ratio of tap-ins to tap-outs is more or less the same for the weekday and weekend averages and thus the loss is systematic.

The most heavily-used stations – from the standpoint of entry and exit activity – in the system are, quite predictably, the network hubs that are either: a) highly central in terms of both network topology and urban geography; or b) 'mainline' stations where suburban commuters transfer from heavy rail to the Tube. The top five stations for tap-ins are Oxford Circus with half a million, followed by Stratford 0.44m, Victoria 0.38m, Canary Wharf 0.33m and London Bridge at 0.33 passengers for the week. The top five stations in terms of tap-outs are Victoria with 0.62m, London Bridge 0.57m, Oxford Circus 0.56m, Liverpool Street 0.43m, and Stratford 0.4m, and this provides a quick picture that the main hubs are determined principally by commuting. In fact, the dominance of the transfer-hubs is likely to be even greater than shown in the Oyster data since many longer-distance commuters were still being issued with traceless magnetic tickets in 2012 and so they are invisible to this analysis.

To break this down into a daily profile, we aggregated entry and exit activity based solely on time-of-day into 96 15-minute intervals and threw away the day-of-week information. In light of the subsequent analysis this might seem to be a surprising choice; however, our testing showed that this approach to aggregation had to meaningful impact on our ability to distinguish between different patterns of activity and had the distinct advantage of accentuating dissimilarity since, over the course of a week of activity, minor fluctuations could accumulate into quite marked differences. So if we look at the aggregate daily profile, the biggest volumes occur in the morning peak between 8 and 9 am where there are some 1.65m people tapping in and 2.07m tapping out and between 5:45 and 6:45pm where there are 1.96m people tapping-in and 1.95m people tapping out. Within these periods, the top five tap-in stations for the morning peak are Brixton, Finsbury Park, Stratford, Victoria, and Ealing Broadway; and for the evening peak Oxford Circus, Canary Wharf, Liverpool Street, Bank, and London Bridge. The top tap-out stations for the morning peak are Canary Wharf, Oxford Circus, Victoria, London Bridge, and Liverpool Street and for the evening peak, Victoria, London Bridge, Waterloo, Oxford Circus, and Liverpool Street.

There is a high degree of consistency between the morning and evening peaks in terms of volumes but less so in terms of the key station hubs. The correlation between tap-in and tap-out volumes for stations in the morning peak is a mere 30% but this rises to 73% for the evening peak. It appears that the two peaks are thus quite different – even though they would appear to be similar in the aggregate system profile – and it is entirely possible that the evening peak is more drawn out than the morning peak and this enables a different degree of mixing of passengers. Moreover, it is also to be expected that the system fills up with less ‘conventional’ commuters during the working day and that this too could complicate the picture. This suggests the system is both complex but self-organising though sorting and mixing.

SIZE AND SCALE: SIMILARITIES AND DIFFERENCES IN DYNAMICS

We will now look at the scale of the system in terms of the size of the stations and also the size of the time periods, but there are clearly many ways of looking at co-variation in the data between stations and time periods and, as we implied at the end of the last section, we can also disaggregate all this data into different time periods – weekdays and weekends – and into different spatially contiguous or non-contiguous clusters of stations. In fact, simply from this one table of data we can generate many more types of analysis: if we were to extend this to tap-outs, and thence to flows between stations across time, the analysis would explode in complexity and scope. This problem points to one of the real challenges of data analysis on rich, real-world, operational systems: big data is more than deep enough to drown in and it, of course, possible to find many statistically significant but functionally meaningless correlations if you ‘let the data speak for itself’. Here, we therefore simply seek to give some sense of what is possible, while a fuller analysis of such data is a major research initiative. This particular example reveals primarily the scale of what is required.

We define the number of tap-ins over a 15-minute time period t for any station i as T_{it} and this can best be considered as a table $\{T_{it}, i=1, 2, ..., 268, t=1, 2, ..., 96\}$. We can examine several attributes of the data from this table: the total volume of tap-in activity at a station is given as $T_i = \sum_t T_{it}$; while the number of tap-ins for each time period across all stations is $T_t = \sum_i T_{it}$. The total tap-ins in the system is the sum of these spatial and temporal quantities, that is $T = \sum_t T_t = \sum_i T_i = \sum_i \sum_t T_{it}$ (17,908,628 tap-ins for the week). To give some sense of what the network looks like in practice, in Figure 1 we show the station hubs scaled by their total tap-in volumes.

Our first foray is the scaling of the stations and for this we have ranked the stations by volume such that they are ordered $T_i(1) > T(2)_j > ..... > T_k(269)$ by rank r which we define as $r(=1) > r(=2) > ..... > r(=269)$. In Figure 2 we show a selection of ranked station volumes taking each of the 15-minute periods beginning on the hour and plot 24 different time periods using an equivalent scaling. The system closes down between roughly 1am and 4am although there are tap-ins in *every* interval which we assume are in places where someone unwittingly taps if the station is open, and even if there are no trains, or are the result of system maintenance and testing. Figure 2 reveals classic scaling of the station volumes and this scaling follows the characteristic of competitive urban systems: it is power law like where there are a few stations with very large flows and a larger number of stations with smaller flows. This pattern is clearly repeated throughout the day and, were we to disaggregate to individual days, the weekdays and weekends would show similar patterns. What is interesting however is that the volumes during the day do suggest that the activity begins to draw itself out as the evening approaches with the evening peak flatter and containing large volumes of travellers, illustrating the dominance of the centre; presumably for entertainment and socialising purposes.

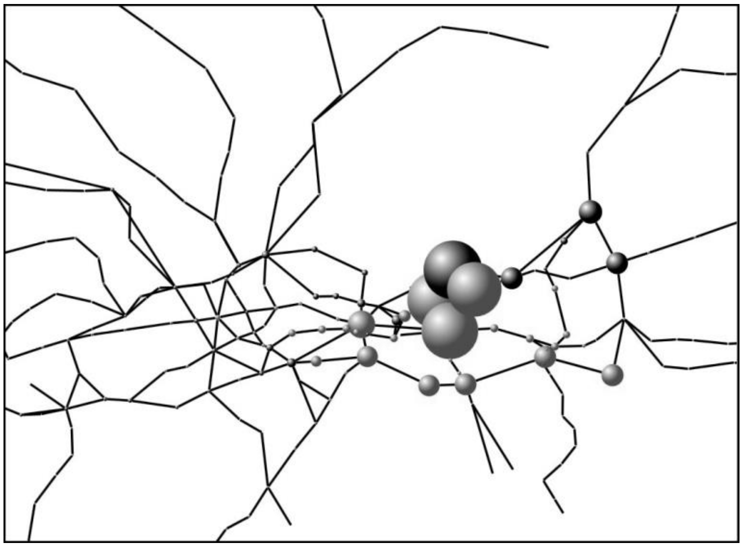

Figure 1. Station Nodes and Underground Links (still to be drawn with correct data)

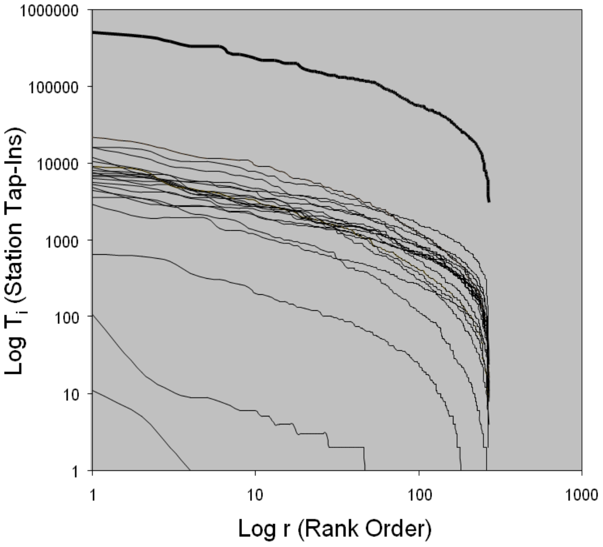

Figure 2. Size and Scaling of Tap-Ins at Stations over a Sample of Time Periods

We can also look at the scaling of the time periods if we rank total temporal flows T_t as $T_t(1) > T(2)_\tau > ..... > T_\delta(96)$ where the subscripts denote the 15 minute periods and the order the rank defined as $r(=1) > r(=2) > ..... > r(=96)$. We could do this for each station but we prefer here to make this simpler by looking at the whole system. In Figure 3(a), we show aggregate tap-in activity over a 24-hour period; these are then shown in rank order in Figure 3(b), showing much less evenness than we would expect from power law behaviour. This feature is further emphasised in the logarithmic *y*-axis used in Figure 3(c), emphasising that a distinct subset of time periods – those which coincide with the peak rush-hour flows – almost divide the system into just two significant time periods: one with massive volumes and one with hardly any volume at all. In plainer-English: the log-log plot highlights the extent to which transit systems operate at or near capacity during only a few key periods (the morning – black – and evening – red – dots on Figure 3).

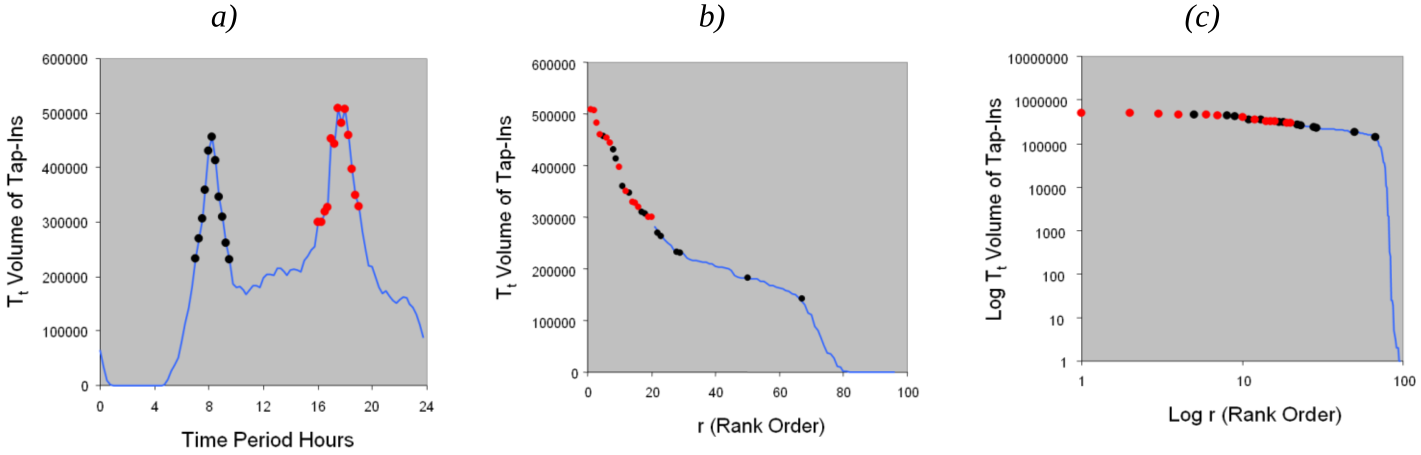

Figure 3. Size and Scaling of Tap-Ins Over Time

The picture that we have presented so far for London is very much in accord with our understanding of behaviour in other large cities with expensive smart card systems rollouts: a strong but very focussed morning peak, illustrating a decanting of population from suburban locations into central work locations, and an even larger evening peak with considerable mixing of movement between the biggest hubs in the core. To make significant progress in this analysis, we need to look at deviations from this picture, but to do so we need to factor the 'routine' of massive peaks and afternoon lulls, and then examine what is remaining. To this end we have devised a measure akin to the location quotient that compares the ratio of a local flow or volume to the global ratio of the same.

The measure involves computing a local quantity – based on the volume of tap-in activity associated with a station during a time period – with respect to the total flow over *all* time periods for that station, and then comparing this to the same ratio calculated for the system as a whole. This is in effect the same idea as a location quotient – which we call here a pseudo-location quotient (ρ) – which measures the relative concentration or dispersion of the local measure relative to the system as a whole. If the measure is greater than 1, this means the local measure is more concentrated than the global while if less than 1, it is more dispersed or de-concentrated than the global. The original location quotient was in introduced by Haig in the Regional Plan for New York in 1928 to measure the relative concentrations or otherwise of different types of industry.

We can write this measure for the flow T_{it} over time periods as:

$$\rho_{it} = \left(\frac{T_{it}}{\sum_t T_{it}} \right) \Bigg/ \left(\frac{\sum_i T_{it}}{\sum_i \sum_t T_{it}} \right) = \left(\frac{T_{it}}{\sum_i T_{it}} \right) \Bigg/ \left(\frac{\sum_t T_{it}}{\sum_i \sum_t T_{it}} \right)$$

Functionally, this is the same as computing the ratio of tap-ins at a station in a time period to all stations over that same time period, and then comparing this to the ratio of flows for that station for all time periods to the total of all stations and time periods. We will compute this measure for all stations and time periods and then examine the relative proportions – relative concentrations – with respect to all stations and all time periods.

There are 296 stations for we can plot our measure ρ_{it}, and we can do the same for time periods with respect to stations. We illustrate these two cases for a sample of TfL stations and time periods in **Error! Reference source not found.**(a) and (b). In short, we compute the measure for each volume T_{it} and then examine their distribution for each station over all time periods or each time period over all stations. If $\rho_{it} > 1$, then the local measure is more concentrated than the global for that station or time period and the reverse is true for $\rho_{it} < 1$. The baseline in the graphs shown in Figure 4(a) and (b) illustrates the situation where the local measure is the same as the system average.

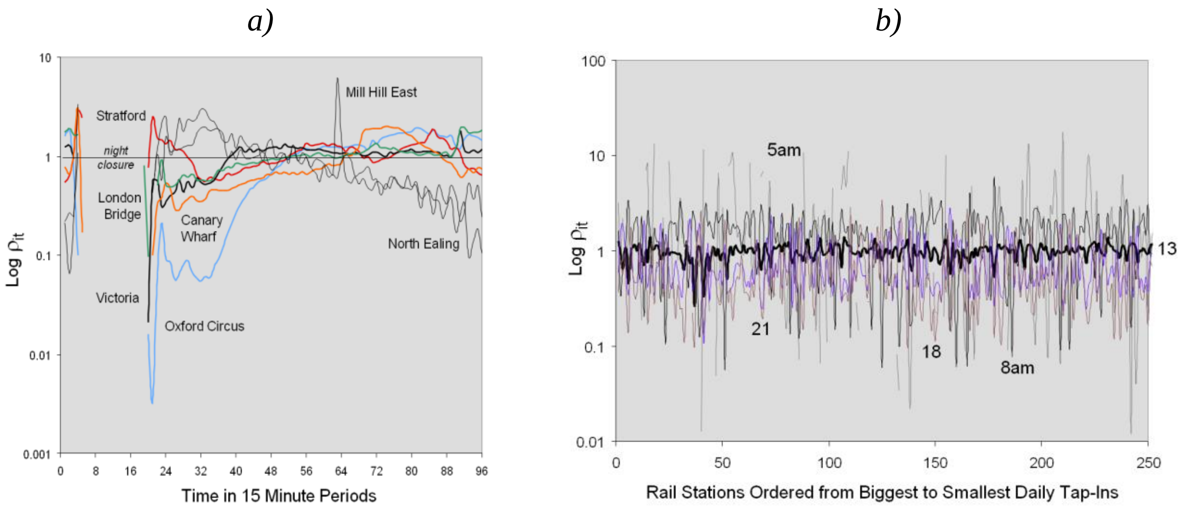

Figure 4. Measures of Difference (ρ_{it}) over Time (a) and Stations (b)

For visual intelligibility, we only plot a selection of these distributions for it we were to plot all 269 stations or 96 time periods, the graphs would be a complete mess. **Error! Reference source not found.**(a) shows the top five, and bottom two, stations by volume, and it is quite clear that the biggest stations become more concentrated as the day wears on whereas the smaller become less so. This is hardly an exhaustive demonstration of this effect but it is systematic and consistent with the notion of the transit system becoming more complex as the day proceeds. In some senses, the system ‘reboots’ in the early morning when it empty of any passengers and the volatility of this is captured by the measure as shown in **Error! Reference source not found.**(b), which reveals that the temporal volatility of stations with respect to concentration and de-concentration is highest in the early morning, stable in the middle of the day, and then becoming more volatile.

To be crystal clear about this pseudo-location quotient, when a station has a coefficient greater than 1, it means that its relative flow is higher than the flow that is suggested by the system average at that time period. If the flow changes and becomes smaller relative to the system average for a time period, this means that it is losing ‘market share’ (Figure 4(a)). The same can be said for a time period which we can graph with respect to all stations. If the time period has a coefficient less than 1, and this changes with respect to the distribution of stations, then this relates to how each station captures more or less flow in terms of what it might be expected where it to perform as the system average (Figure 4(b)). In fact the first interpretation is much more intuitive than the second because it is much easier to think of a station losing or gaining flow over time than it is for a time to be gaining or losing its flow over stations. This is because the dynamics is temporally driven, not spatially at least not in terms of what this data is actually revealing. We can also convolute the pseudo-location quotients even further and order them by their rank and sixe over stations or time periods, just as we did in Figure 2 and 3 for the aggregate data, but these are not very easy to interpret and as such represent the limits to this analysis and we will not illustrate these rank size distributions.

PRINCIPAL COMPONENTS OF THE PSEUDO-LOCATION QUOTIENTS

As any visitor to London can attest, the volume of activity can vary dramatically from station to station, and from one time period to the next. The busiest stations are, not entirely surprisingly, well-known work and leisure locations such as Oxford Circus and Canary Wharf, and major interchanges at mainline stations such as Victoria, Stratford, Liverpool Street, and London Bridge; each of these recorded more than 400,000 tap-ins during the week for which data was available. In contrast, peripheral stations such as Angel Road and Morden South recorded fewer than 500 tap-ins across the entire week! This tap-in hierarchy is, broadly, preserved across time as well: Canary Wharf and Oxford Circus also have the busiest times-of-day, experiencing peaks during the afternoon rush hour of more than 20,000 entry taps in a 15-minute period.

Of course, in any given time period, the largest stations might *not* actually be the busiest in either absolute or relative terms: we would naturally presume that earlier peaks in activity would correlate with inward commuter flows, while later ones would link to commuter outflows and to nightlife activities. In addition, we can also consider whether the intensity of these peaks as a function of overall flows sheds light on trip purpose: we might assume that an 8am peak that represents 75% of daily tap-ins maps on to a relative dominance of commuters and absence of other land uses. Our working hypothesis is therefore that the timing and intensity of each station’s peaks and troughs in tap-in activity captures something of the spatial structure of the city because behaviour will vary with location and land-use.

However, in order to compare data from stations where the maximum inflow is measured in the dozens with stations where it is measured in the tens of thousands, we need a way of standardising across a range of tap-in levels. A simple approach would be range standardisation, in which we simply rescale the interval data so that the maximum value (20,000+ tap-ins at Oxford Circus, say) becomes 1 and all other values are scaled relative to this. Unfortunately, this approach will cause subtle analytical issues down the line because it fails to grapple with the changing volume of flows across the network as whole (*e.g.* that the evening rush hour has a 'shoulder' not visible in the morning rush hour but actually has higher peak volumes!).

Use of the Location Quotient to Standardise Tap-In Activity

As discussed above, we can condition our expectation of activity at a station by the level of activity in the system as a whole using the pseudo-location quotient. If there were *no meaningful spatial or structural difference* between stations, then as the total volume of tap-ins increased around rush hour, all stations would increase at a similar rate and to a similar degree. To put this another way: with the pseudo-location quotient, we start from the *naïve expectation* that all stations have a constant share of total usage and standardise the data with respect to that expectation; we obviously know that this will not be the case, but it is the deviations from this naïve expectation that are of the most interest to us as a tool for classification.

So the pseudo-location quotient $\{\rho_{it}\}$ standardises the data such that unity – a value of 1 – means that the volume of tap-ins at station i in time t is the same as the share of tap-ins for i across the entire week. Figure 4 gives a sense of how this standardisation process impacts the pattern observed for stations selected on the basis of very different signature patterns of overall activity. The two enormous evening peaks are for Canary Wharf and Oxford Circus, highlighting how even peaks with similar timings and magnitudes can nonetheless express quite different behaviours and the degree to which the $\{\rho_{it}\}$ can capture that variation.

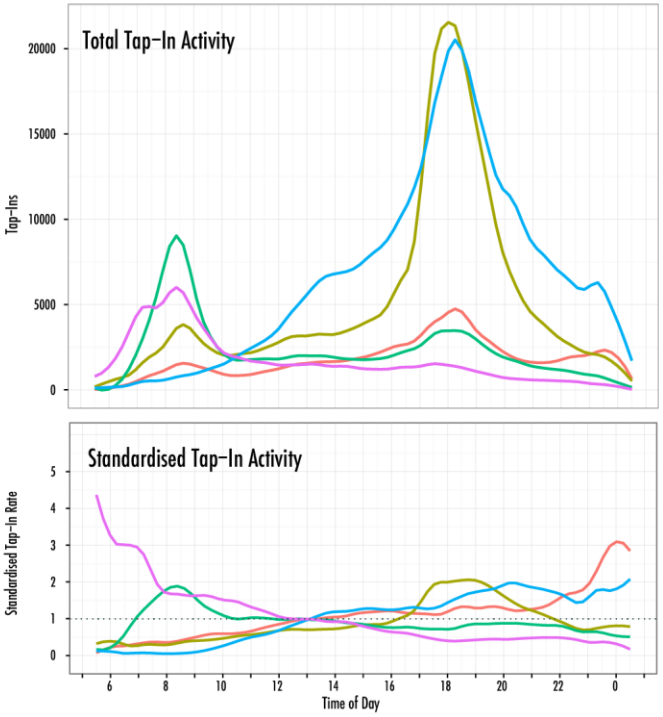

Figure 4. Tap-In Activity over Time at Selected Stations

Use of Principal Components Analysis as a Classification Mechanism

Having standardised the data to control for the variation in activity levels at each station, we are now in a position to try to classify stations into clusters by looking at the degree to which they display similar – or divergent – patterns over time. This is a kind of dimensionality reduction analysis in which we want to reduce the complex, time-varying pattern into a single number, or set of numbers, that give us a similarity measure between stations. One common way to achieve this is the use of Principal Components Analysis (PCA), which yields a set of ordered vectors, each of which captures some common aspect of the temporal pattern seen above while also providing a single metric that tells us how much of that pattern is observed at any given station. In order to prevent peaks in the standardised data – which typically coincide with periods of much lower overall usage and, consequently, greater instability – overwhelming the subsequent clustering process, we perform PCA on column-normalised data, meaning that each time period has *now* been rescaled to the range between 0 and 1. Figure 5 summarises the output of the PCA process, giving a sense of how much each dimension contributes to the standardised activity levels and of what each dimension 'looks like' over time: the more important vectors (*i.e.* the lower-numbered dimensions on the left) also have more structure (*i.e.* on the right), with the smaller ones looking increasingly like noise. The results indicate that the majority of the system can be described by just 14 eigenvectors (Reades et al. 2009)., each of which accounts for at least 1% of the observed variation, with the first five explaining 67.6% of the total

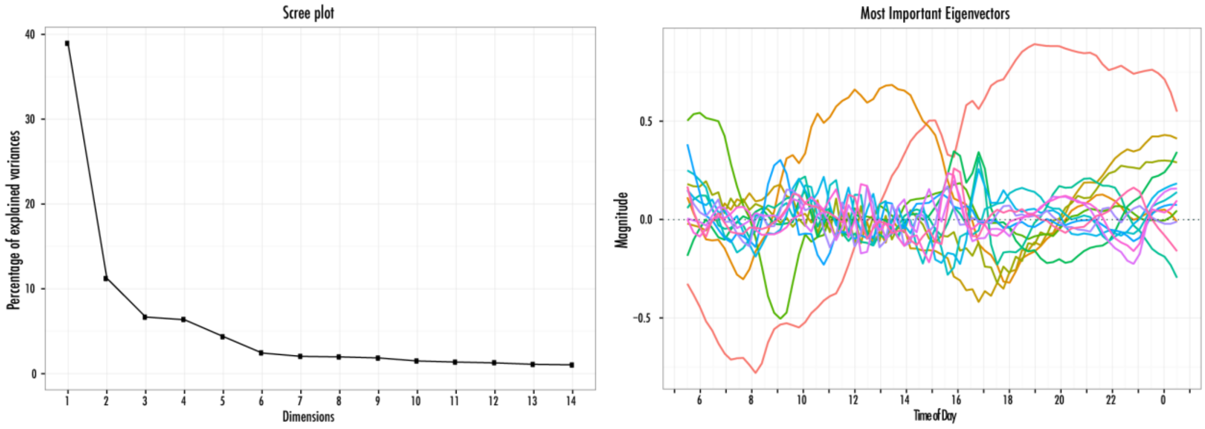

Figure 5. The Principal Components Analysis of Tap-In Activity

From this it can be seen that the key periods in terms of a spatio-temporal analysis of the network are an extended evening peak running from 5pm to midnight, an early morning peak between 5:30 and 8am, and a later morning peak running from 8:30am to 4:30pm. Broadly, the eigenvector plot also suggests that the evening variation across stations is greater than in the morning since the spread of the standardised values is greater, and that a set of stations exists where late-night activity levels is relatively higher than the evening rush-hour. Figure 5 is therefore largely intended to provide a more intuitive understanding of the behaviour captured in each derived dimension.

Using Clustering on PCA Output

We turn now to the eigenvalues in order to perform a classification process based on the extent to which each station expresses one or more of the behaviours briefly discussed above (Calabrese et al., 2010). However, it is worth reflecting further on what it is that we are attempting to capture:

PCA represents the mapping of input data on to a new set of (orthogonal) dimensions that maximises the variance along each derived axis in order of descending ‘importance’. Implicitly, the eigenvalues represent the location of a station along each axis and, consequently, stations that behave differently should end up quite ‘far’ from each other in the derived data space of 14 dimensions, while those that behave in similar ways should end up quite ‘near’ to each other in this multi-dimensional space.

In other words, we are seeking to measure distance between station pairs and to use this as our clustering criterion. So although a wide range of clustering algorithms exist, with *k*-means being by far the most common, the most appropriate choice here is therefore the PAM (Partitioning Around Medoids) which uses as the key criterion for assignment, the sum of the distances between an observation and all the other members of the cluster to be at a minimum. As with *k*-means, PAM requires that the number of clusters be specified in advance; however, the data set is small enough that it is relatively trivial to test a wide range of possible cluster counts while searching for the most effective partitioning using the ‘silhouette’ measure.

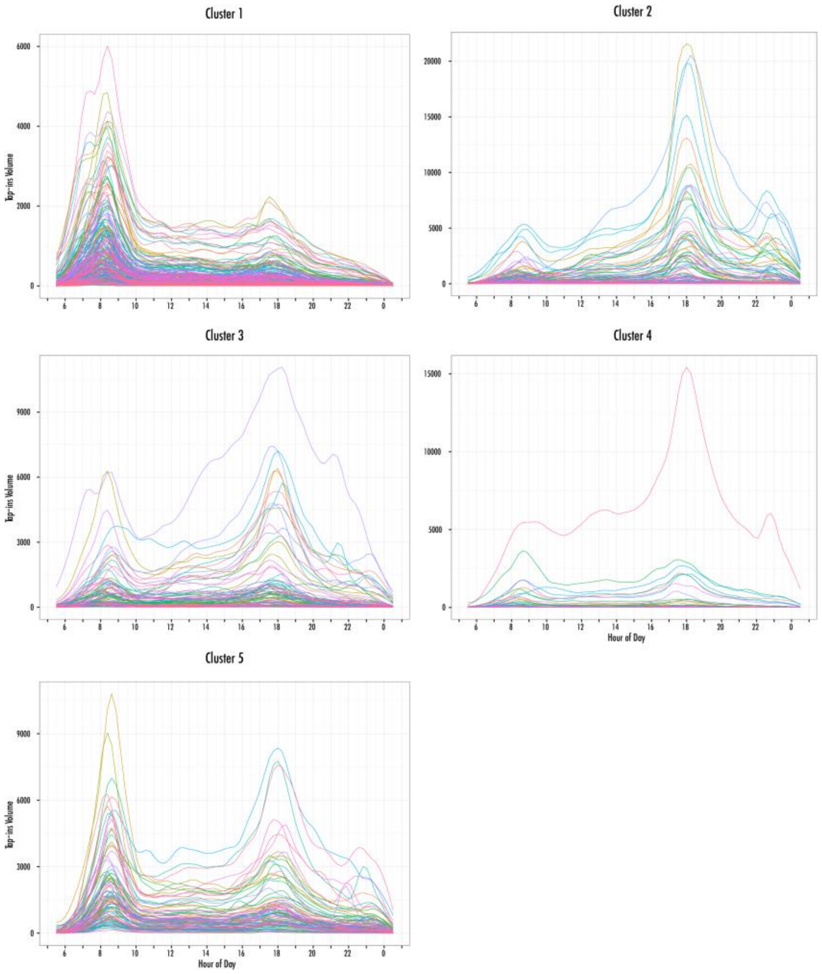

Figure 6. Tap-In Values by Station and Cluster Across the Time Periods

Results from the iterative clustering process which we show in Figure 7 indicate that 5 clusters yield the highest values – indicating that the clusters are clearly distinguishable – at a scale that is still meaningful for interpretation since it would be extraordinarily difficult to draw useful insights from an analysis that yielded 15 or 20 clusters. Recall that the clustering was applied to standardised *and* normalised data, and as such, these clusters contain stations whose contributions to the system across a representative 24-hour period are broadly similar – they will *not* line up at each and every period, but at periods of high- and low-standardised activity, they should be similar. Now that the clustering is complete, however, we can take the station classification and go back to the *original* data to evaluate what has been picked up by the decomposition and partitioning processes.

Finally, of course, we can also map the stations in order to assess the extent to which meaningful spatial structure has been derived from the station tap-ins. These are shown in Figure 8.

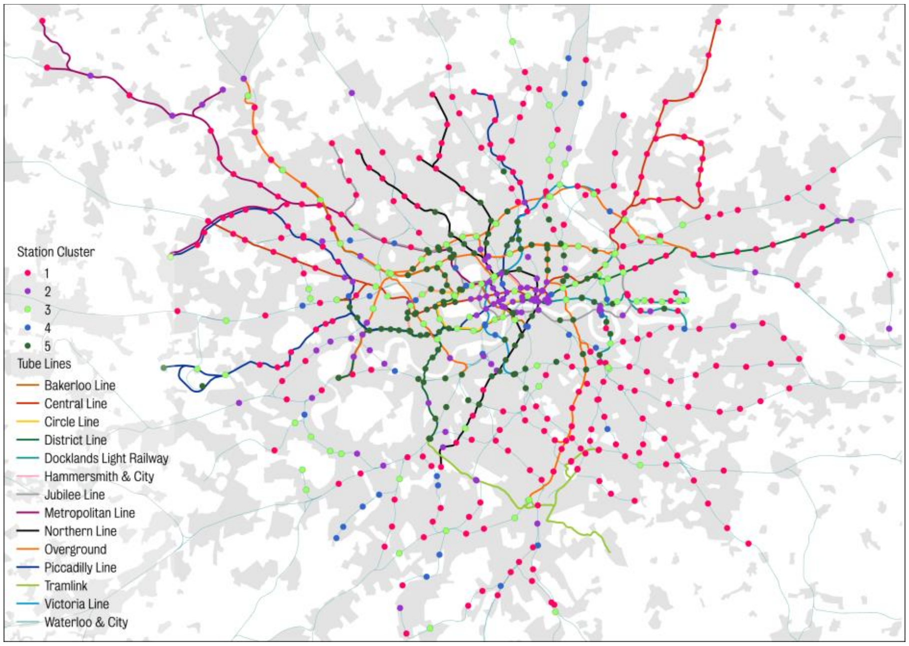

Figure 7: Cluster Map of the Stations From the Clustering the Principal Components

EXTENDING THE ANALYSIS: VARAIABILITY AND DISRUPTION

xxxxxxx

REFERENCES

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tir, M., and Zook, M. (2015) Everyday Space–Time Geographies: Using Mobile Phone-Based Sensor Data to Monitor Urban Activity in Harbin, Paris, And Tallinn, **International Journal of Geographical Information Science 29** (11), 2017-2039.

Algueró, S. (2015) **Using Smart Card Technologies to Measure Public Transport Performance: Data Capture and Analysis**, Industrial Engineering,

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Machowicz, M., Ouzounis, G., and Portugali, Y. (2012) Smart Cities of the Future, **European Physical Journal Special Topic**, **214,** pp. 481-518.

Calabrese, F., Reades, J. and Ratti, C. (2010) Eigenplaces: Segmenting Space Through Digital Signatures, **IEEE Pervasive Computing**, **9** (1), 78-84.

Gong, Y., Y. Liu, Y. Lin, J. Yang, Z. Duan and G. Li (2012) Exploring spatiotemporal characteristics of intra-urban trips using metro smartcard records. **Geoinformatics** (GEOINFORMATICS), 2012 20th International Conference on, IEEE.

Gonzalez, M. C., Hidalgo, C. A., Barabasi, A. L. (2008) Understanding Individual Human Mobility Patterns, **Nature 453** (7196), 779-782.

Gordillo, F. (2006) The Value of Automated Fare Collection Data for Transit Planning : An Example of Rail Transit OD Matrix Estimation, Dept. of Civil and Environmental Engineering, MIT, Cambridge, MA, http://hdl.handle.net/1721.1/38570

Gordon, J. B., H. N. Koutsopoulos, N. H. Wilson and J. P. Attanucci (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.

Long, Y. and J.-C. Thill (2015) Combining Smart Card Data and Household Travel Survey to Analyze Jobs–Housing Relationships in Beijing, **Computers, Environment and Urban Systems 53**, 19-35.

Manley, E., Zhong, C., and Batty, M. (2016) Spatiotemporal Variation in Travel Regularity through Transit User Profiling, submitted to **Transportation**, under review

Morency, C., Trépanier, M., Agard, B. (2006) Analysing the Variability of Transit Users Behaviour with Smart Card Data, **Intelligent Transportation Systems Conference** ITSC'06 IEEE, 44-49.

Munizaga, M. A. and Palma, C. (2012) Estimation of a Disaggregate Multimodal Public Transport Origin-Destination Matrix from Passive Smart Card Data From Santiago, Chile, **Transportation Research 24C,** (12), 9-18.

Munizaga, M. A., Devillaine, F., Navarrete, C. , and Silva, D. (2014) Validating Travel Behavior Estimated from Smartcard Data, **Transportation Research 44C**, 70-79.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C (2012) Correction: A Tale of Many Cities: Universal Patterns in Human Urban Mobility, **PLoS ONE 7(5):** e37027.doi:10.1371/journal.pone.0037027

Pelletier, M. P., Trépanier, M., and Morency, C. (2011) Smart Card Data Use in Public Transit: A Literature Review, **Transportation Research** Part C: Emerging Technologies **19**, 557-568.

Reades, J., Calabrese, F., and Ratti, C. (2009) Eigenplaces: Analysing Cities Using The Space-Time Structure of the Mobile Phone Network, **Environment and Planning B**, **36**, 824-836.

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

Silva, R., Kang, S. M., and Airoldic, E. M. (2015) Predicting Traffic Volumes and Estimating the Effects of Shocks in Massive Transportation Systems, **Proceedings of the National Academy of Sciences**, **112**, 18, 5643–5648.

Williams, M. J., and Musolesi, M. (2016) Spatio-Temporal Networks: Reachability, Centrality and Robustness. **Royal Society Open Science**, 3: 160196, http://dx.doi.org/10.1098/rsos.160196

Zhong, C., Arisona, S. M., Huang, X., Batty, M., Schmitt, G. (2015) Detecting the Dynamics of Urban Structure through Spatial Network Analysis, **International Journal of Geographical Information Science 28**, 2178-2199.

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., and Schmitt, G. (2016) Variability in Regularity: A Comparative Study of Urban Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data, **PLoS ONE 11**(2): e0149222. doi:10.1371/journal. pone.0149222