Anti-gravity Underground?

Chris Smith[1], Daniele Quercia[2], and Licia Capra[1]

[1] Dept. of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK {C.Smith|L.Capra}@cs.ucl.ac.uk
[2] Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK dq209@cl.cam.ac.uk

Abstract. Since its introduction in 1946, the Gravity Model has proven to be a valuable aid in understanding and predicting forces at play in human mobility. Historically, the model has been applied primarily at the intercity and interstate level. In this paper we investigate whether the scope of the model extends to the intra-urban. Using a large dataset of trips made on the London rail network, we derive a number of proxies for mass and distance, and using these proxies we examine the extent to which the estimated passenger flows of the Gravity Model fit the observed data. We find that there is a good correlation between the estimates and observations, with a Pearson's correlation coefficient (PCC) of 0.720. We then extend the investigation to examine how well the model fits when focusing on trips of differing purpose, and we also modify the model to take into account the attractiveness of a location. We find that the original model better fits leisure trips than commute trips, but that the modified model closes this gap, and indeed offers an improvement overall, increasing PCC up to 0.789. We close with a discussion of the validity of our results and the direction of future research.

1 Introduction

The share of the world's population living in cities has recently surpassed 50%, and it is expected that by 2025 another 1.2 billion people will be living in urban areas. Municipal planners will face an increasingly urbanised and polluted world, with cities everywhere suffering an overly stressed road transportation network. Building effective public transport systems, capable of absorbing the increasing load, has thus become an urgent priority, both to provide a good quality of life and a cleaner environment, and to remain economically attractive to prospective investors and employees. To make effective large capital investments concerning the public transport system, urban transportation analysts and planners will need to understand and predict citizens' flows *within* the city.

To understand and predict population flows, researchers have proposed various spatial interaction models over the years. The most widely used one so far has been the Gravity Model [14]. In analogy to the gravitational interaction between planetary bodies, the model posits that the interaction between two places (e.g., two cities) is proportional to their mass (e.g., their population) and inversely proportional to their distance. The model has been successfully used to

describe 'macro-scale' interactions (e.g., between cities, and across states), using both road and airline networks (e.g., [6], [2]). In this paper, we are interested in investigating whether this model can be used at 'micro-scale' level too, to capture citizens' movement *within* a city, travelling via mass public transport systems.

In this work, we use anonymised data collected by Transport for London (TfL) across the entire London rail network (Section 3), to derive a Gravity Model of passengers' flow that captures intra-city dynamics at station level (Section 4). We then observe how well the model fits the observed data, both when looking at the whole set of recorded journeys, and when analysing urban trips of different purpose separately (Section 5). We find that, in its present formulation, the model is accurate in capturing passengers' flow for leisure trips, but less so for work commutes. We then introduce a modified version of the model which incorporates a notion of the attractiveness of a station. The modified version offers an improvement, yet the variation in the results we observe suggests that passenger flows respond to other socio-economic forces that we have not yet captured within the model; we thus conclude with a discussion of what information we plan to integrate next in the model (Section 6).

2 Background and Related Work

In its simplest formulation, the Gravity Model [14] states that the interaction $T_{i,j}$ between two places i and j is proportional to the product of their populations P_i and P_j over their distance $d_{i,j}$:

$$T_{i,j} = k\frac{P_i^{\alpha} P_j^{\beta}}{d_{i,j}^{\gamma}}, \tag{1}$$

where the exponents α, β, γ and the scaling factor k are adjustable parameters, practically chosen so to fit the empirical data being modelled. Over the years, this simple model has been expanded and refined in a variety of ways, and its application has gone well beyond the transportation domain [6, 2]. For example, it has been used to model inter-city phone calls [8], the spreading of infectious diseases [1], and cargo ship movements [7]. Despite some criticism [13], partly related to the lack of a rigorous derivation of its parameters, this highly generalised spatial interaction model continues to be widely used in practice, if not for prediction at least for understanding complex network dynamics.

An area where this model has been little investigated is the *intra*-city transport domain. As urban cities are fast growing in number and size, understanding the complex dynamics that govern the use of an often diverse multi-modal public transport infrastructure is of high importance, for both transport providers and urban planners. To date, the major hindrance to this type of investigation has been the lack of available datasets that accurately capture citizens' movement within the city. However, as Automated Fare Collection (AFC) systems are deployed in cities worldwide, continuous and fine-grained records of citizens' movement within the urban public transport infrastructure are becoming

available. These datasets have so far been investigated, for example, for demand modelling [4], station transfer analysis [5], service reliability measurements and average route travel time estimation [3].

In this work, we use AFC data to test whether the Gravity Model can be used to explain and possibly predict passengers' flows *within* a metropolitan city. Our investigation expands on the studies of the Gravity Model in transport networks conducted so far in two ways: *granularity*, as 'places' in our model are as fine-grained as underground and train stations within the city of London, UK; and *scale*, as we analyse over 500 stations (i.e., the entire TfL rail network). Next, we describe the dataset at hand.

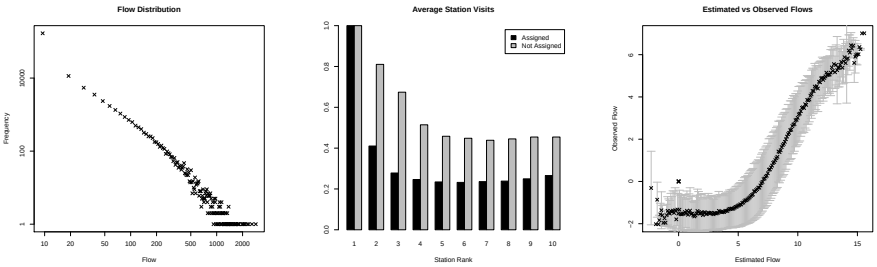
3 The London Public Transport Network and Dataset

The public transport system in London consists of several interconnected sub-systems, incorporating multiple modes of transport. These include, but are not limited to, the London Underground (known colloquially as the Tube), the Overground rail system, an extensive bus network, water-borne transport and parts of the UK National Rail network, of which many services terminate in London. For the purpose of this study, we focused on the rail sub-networks, comprising the Tube, the Underground and the UK National Rail stations and tracks within the Greater London area, for a total of 588 stations.

In 2003, Transport for London (TfL) introduced an RFID-based technology, known as Oyster card, which at the present time accounts for 84.5% of all journeys made in the London public transport system, with the rest made using traditional paper based magnetic stripe tickets[1]. The dataset we use consists of a record of every journey taken on the London rail network using an Oyster card, in the 31 days of March 2010. A record in the dataset is a tuple of the form: $\langle u, (o, d), t_o, t_d \rangle$, recording that an anonymised user id u travelled from station o at time t_o, to station d at time t_d. In total the dataset contains 76.6 million journeys.

Using these records as a basis to investigate the applicability of the Gravity Model to transportation within a city has both advantages and disadvantages with respect to approaches based on survey data (e.g, [11] based on the 2001 UK census, and [13] which uses the 2000 US census). Being expensive to process and analyse, survey data is only collected periodically (e.g, every 10 years in the UK); furthermore, it contains information about respondents' travel habits at a coarse granularity (e.g., the city where they live/work). In contrast, AFC data provides a detailed, accurate picture of the usage of the urban public transport system in real time. On the other hand, survey data usually contains richer semantic information than AFC data, including home and work location of each respondent (though changes happening within the 10-year data collection cycle are lost), as well as purpose of travel. In the following sections, we describe how we have mined public transit records to infer this semantic information and construct the Gravity Model around it.

[1] http://www.whatdotheyknow.com/request/oyster_card_usage - Retrieved 9/03/12



(a) Flow Distribution (b) Station Visits (c) Prediction results

Fig. 1. (a) Average number of daily journeys between pairs of stations, (b) mean relative distribution of station touch-ins in am-peak, and (c) Newton gravity model's estimated flows vs. observed flows.

4 Building an Underground Gravity Model

We derive an instance of the Gravity Model for the London rail network in three steps: (1) first, we derive a proxy for mass (population) P_i for each station i in the dataset; (2) second, we derive a proxy for distance $d_{i,j}$ between each pair of stations; (3) finally, we establish actual (empirical) interactions $T_{i,j}$ against which to test the accuracy of the model. In the remainder of this paper, we refer to a specific instance of the Gravity Model, commonly referred to as the Newton Model, whereby we set $\alpha = \beta = k = 1$, and $\gamma = 2$ (as done, for example, in [6]). The simplified model thus states:

$$T_{i,j} = \frac{P_i P_j}{d_{i,j}^2}. \tag{2}$$

(1) Proxy for Mass $\mathbf{P_i}$. In transportation networks, mass is often represented as population size (e.g., [14, 6]). Survey data is used to accurately associate every citizen to their home location. In our case however, each station is represented by a single point and as such, population is undefined. We therefore need to derive a proxy for population P_i at each station i. We tested two: *population density*, and a second proxy derived from the transit data as follows. We process all travel records found in our dataset on a per user u basis; we then rank departure stations o based on their popularity (i.e., how often user u has 'touched in' at station o). In order to distinguish genuine London residents from occasional visitors, we apply the following restriction: we only count departures within the morning peak period, 6:30am to 9:30am, based on the assumption that the vast majority of journeys in this period will be commutes from u's home to a place of work. In so doing, we also avoid counting departures from u's other frequented stations, such as work place in the evening. The downside is that we may exclude residents whose main use of the rail network is not for commuting. For every user u we thus compute a ranking vector $R_u = [r_1, \ldots, r_N]$, where $N = 588$ is the total number of stations in the dataset, and r_k is the number of times u

has departed from station o_k (with $o_{k=1}$ being the most frequently visited origin station). We then assign users a home station according to the following set of rules that we apply in sequence: (a) if $r_1 \leq 2$ (the user's most visited origin station has been visited no more than twice in a whole month), the user is not assigned a home station; (b) if $r_1/r_2 > 0.5$, assign o_1 as home station; (c) if $r_1/r_2 \leq 0.5$ and $r_2/r_3 > 0.5$, assign both $o_1$and o_2 as home stations; (d) otherwise, the user is not assigned a home station. Note that up to two stations can be designated as a home station for a user, since in some parts of London there may be more than one station within equal distance from a user's residence, and the choice of which one to depart from may depend on factors which vary day to day. Finally, our second proxy for population P_i is the total number of users assigned to station i; note that not all users contribute to station populations; in fact, the above steps discard 76% of users whose travel records do not reveal any preferential origin station (case (d)). Figure 1(b) shows, seperately for users included and not included in any station's population, the average relative number of touch-ins at their top ten stations. We see that included users touch-in at their third most popular station just 28% of the number of times they visit their most popular, compared to 67% for users not included.

(2) Proxy for Distance $\mathbf{d_{i,j}}$. We tested the gravity model using the euclidean distance between stations, as commonly done in the literature; however, *geographically near* stations may be quite distant in terms of *path length* in the rail network, and indeed this measurement performed poorly. To overcome this, we defined several other proxies for distance: *number of hops* between stations, with a penalty of 2 hops added for each interchange; *mean transit time* between stations computed from all instances of trips from station i to station j $(t_j - t_i)$ recorded in the dataset; *ranked euclidian distance*; and *ranked mean transit time* (ranked distance has been shown to work well in [10] and [13]).

(3) Empirical Interactions $\mathbf{T_{i,j}}$. Lastly, in order to quantify how well the Gravity Model estimates actual intra-city flows, we establish the ground truth $T_{i,j}$, that is, the flow between stations i and j, simply as the average number of *daily* journeys between every pair of stations, as recorded in our dataset. Figure 1(a) shows the distribution of average number of daily journeys, which follows a power law. By then building a graph whose nodes are stations and edge weights are average *daily* journeys (flow) between station pairs, we find that the graph is highly connected, with 59% of all possible edges present.

5 Testing the Model

Based on the proxies for mass and distance described in the previous section, we tested the accuracy of the Gravity Model (Equation 2) in estimating passenger load on the London rail network. To quantify the overall accuracy of the model with different values for mass and distance, we compute the Pearson's correlation coefficient between observed and estimated flows (note that we take the logarithm of each value in order to account for the skewed distribution of flows; for all correlation coefficients presented, the p-value is effectively 0). Results are

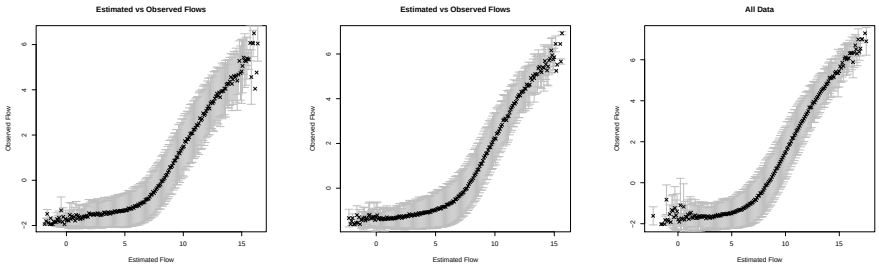
presented in Table 1. The best performing combination is that of P_i = number of home users assigned to i, and $d_{i,j}$ = mean travel time between i and j. Figure 1(c) shows observed vs. estimated edge weights (i.e., interaction values $T_{i,j}$) for this combination. The edges are first binned by estimated flow, then we plot the mean estimated flow in each bin vs the mean observed flow of the edges in each bin. The error bars show the standard deviation of the observed flows in each bin. Using the proxies we have defined, the gravity model provides estimates which correlate fairly well with the observed flows.

So far in our analysis, we have not differentiated between trip purposes but rather looked at the whole dataset at once. When studying population movement at inter-city or inter-state level, studies have often focused on single-purpose journeys (typically the work commute). We thus repeat our analysis (using our best performing proxies) focusing on two separate slices of the whole journey dataset, which we refer to as *commute trips* and *leisure trips*. In terms of ground truth flow distribution $T_{i,j}$, the former comprises of all journeys made on weekdays during the 6.30-9.30am peak period; the latter comprises all journeys made on weekends at any time. Previous research into system-wide activity levels has shown that the am-peak period is dominated by commute trips, and weekends lack the same acitivity profile [9], indicating that the vast majority of trips are for leisure purposes, broadly defined. Table 1 also shows the correlations between observed flows from each data slice and the estimated flows from the model. The results suggest that the model better predicts passenger flows for leisure trips than for commute trips.

When looking at different trip purposes, the directionality of the flow may play an important role in estimating spatial interactions in a city. In order to capture the directionality of flows, we define the *attraction* A_j of station j using the same procedure described in Section 4(1) to assign users to home stations; this time, rather than counting users' touch-ins, we focus on users' touch-outs instead. In so doing, we implicitly reflect the attractiveness of a station for work purposes in the am-peak period, for leisure purposes in the weekend period, and any purpose for the complete data slice. We then replace P_j in Equation 2 with A_j and recalculate the estimated trip flows. The results from our modified version of the Gravity Model, for commute, leisure and all trips are shown in Figures 2(a), 2(b) and 2(c) respectively. The Pearson's correlation coefficients are higher than when using the symetrical model: 0.743 for both work and leisure

Data Slice	P_i **Proxy**	**Euclidian**	**Hops**	**Time**	**Ranked Eucl.**	**Ranked Time**
Complete	Pop. Density	0.315	0.401	0.492	0.292	0.416
Complete	Home Users	0.554	0.700	0.720	0.554	0.675
Am Peak	Home Users	0.418	0.523	0.565	0.462	0.635
Weekend	Home Users	0.520	0.619	0.667	0.532	0.630

Table 1. Comparison of test results using different proxies for mass and distance, and for each data slice.



(a) Commute Trips (b) Leisure Trips (c) All Trips

Fig. 2. Unidirectional Gravity Model estimated flows vs. observed flows for each data slice.

trips, and 0.789 for all trips. This suggests that the Gravity Model seems to perform better if trip directionality is taken into account, that is, when the mass of the destination station reflects its *attractiveness*. This is particularly true of commute and leisure trips, where we see the greatest improvement over the symmetrical model. It is our plan to continue in this line of enquiry, and investigate how socio-economic factors (e.g., crime rates, employment rates, and income) in different areas of London can be used to better represent the distance (deterrence) function $d_{i,j}$.

6 Discussion and Future Work

In this position paper, we have presented exploratory work into how AFC datasets, capturing citizens' movement within urban mass transport systems, can be mined and used to examine the validity of well-established theories of spatial interactions at an unprecedented level of granularity and scale. In particular, we find that the Newton version of the Gravity Model, with transit time used as proxy for distance, and with traveller's population at origin stations used as proxy for mass, offers a reasonable estimation of passengers' flows, but performs less accurately when trying to model leisure trips, and yet even worse for commute trips. We then introduced a uni-directional version of the model using a new proxy to capture the attractiveness of a station. We found that this offered a marked improvement over the gravity model. Despite the correlation between estimated and observed flows being overall high, the variation we observed in the results suggests that there are other forces at play in determining urban movement via public transport, that have yet to be accounted for in the model.

We are continuing this line of investigation, looking at which additional socio-economic factors may be used to unveil the complex dynamics behind urban flows, and how we can integrate them with the model as, for example, alternative proxies for mass and distance. The factors we are looking at come from

two main sources: publicly-available government-maintained datasets[2], offering information such as crime rates, employment rates, income, and education levels of different areas of London; and social-media datasets (e.g., Twitter, that can be mined to reveal happiness of citizens in different parts of the city [12], and Foursquare check-ins, that reveal attractiveness of an area), that offer a more dynamic view of the pulse of the city.

Acknowledgements. The research leading to these results has received funding from the European Community under Grant Agreement n. 234239.

References

1. Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, December 2009.
2. Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *PNAS*, 101(11):3747–3752, 2003.
3. J. Chan. Rail Transit OD Matrix Estimation and Journey Time Reliability Metrics Using Automated Fare Data. Master's thesis, MIT, Department of Civil and Environmental Engineering, June 2007.
4. K. Chu and R. Chapleau. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *J. of the Transportation Research Board*, 2063, 2008.
5. W. Jang. Travel Time and Transfer Analysis Using Transit Smart Card Data. *Journal of the Transportation Research Board*, (3859), 2010.
6. Woo-Sung Jung, Fengzhong Wang, and H. Eugene Stanley. Gravity model in the korean highway. *EPL (Europhysics Letters)*, 81(4):48005, 2008.
7. Pablo Kaluza, Andrea Kölzsch, Michael T. Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of The Royal Society Interface*, 7(48):1093–1103, July 2010.
8. G Krings, F Calabrese, C Ratti, and V D Blondel. Urban gravity: a model for intercity telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, (07), 2009.
9. Neal Lathia, Jon Froehlich, and Licia Capra. Mining public transport usage for personalised intelligent transport systems. *Data Mining, IEEE International Conference on*, 0:887–892, 2010.
10. David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.
11. A. P. Masucci and G. J. Rodgers. The network of commuters in london. *Physica A: Statistical Mechanics and its Applications*, 387(14):3781–3788, June 2008.
12. Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. Tracking gross community happiness from tweets. In *Proc. of the ACM Conference on Computer Supported Cooperative Work*, pages 965–968, 2012.
13. Filippo Simini, Marta C. Gonzalez, Amos Maritan, and Albert-Laszlo Barabasi. A universal model for mobility and migration patterns. *Nature*, February 2012.
14. George Kingsley Zipf. The p1 p2/d hypothesis: On the intercity movement of persons. *American Sociological Review*, 11(6):pp. 677–686, 1946.

[2] For London: http://data.london.gov.uk/