Finger On The Pulse: Identifying Deprivation Using Transit Flow Analysis

Chris Smith
University College London
London, UK
chris.smith@ucl.ac.uk

Daniele Quercia*
Yahoo! Research
Barcelona, Spain
dquercia@yahoo-inc.com

Licia Capra
University College London
London, UK
l.capra@ucl.ac.uk

ABSTRACT

A common metaphor to describe the movement of people within a city is that of blood flowing through the veins of a living organism. We often speak of the 'pulse of the city' when referring to flow patterns we observe. Here we extend this metaphor by hypothesising that by monitoring the flow of people through a city we can assess the city's health, as a nurse takes a patient's heart-rate and blood pressure during a routine health check. Using an automated fare collection dataset of journeys made on the London rail system, we build a classification model that identifies areas of high deprivation as measured by the *Indices of Multiple Deprivation*, and achieve a precision, sensitivity and specificity of 0.805, 0.733 and 0.810, respectively. We conclude with a discussion of the potential benefits this work provides to city planning, policymaking, and citizen engagement initiatives.

Author Keywords

public transport; urban computing; well being; data mining

ACM Classification Keywords

H.1.2 User/Machine Systems: Human information processing; H.5.m Information Interfaces and Presentation: Miscellaneous

INTRODUCTION

Large contemporary cities exhibit complex spatial and temporal dynamics, such that quality of life and the environment can vary significantly between adjacent areas and change over relatively short time periods. It is well known that the planet is undergoing a rapid population shift towards urban environments, with an estimated growth of 5 million new city dwellers each month in developing countries. With rapid growth comes an ever increasing need for effective planning and management of urban infrastructure. Indeed, a 2008 United Nations report into the state of the world's cities showed that inequality in urban environments is on the rise, as some areas and communities benefit more than others from economic growth and investment in public services [20]. The report states that urban inequality has a detrimental effect on citizens' health, education and participation in society and the economy, which in turn leads to social unrest and the diversion of resources from productive public investment toward security services, thus exacerbating the problem further. Therefore the importance of averting and rebalancing such inequality cannot be overstated.

To efficiently allocate limited resources, policymakers and agencies first need to identify which areas are in most need of intervention in order to alleviate deprivation. Assessment of city neighbourhoods' relative prosperity or deprivation may take place only once every few years, and the larger the assessment window, the more likely that problem areas will deteriorate. Thus, developing new methods of identifying urban inequality swiftly and at low cost would offer significant social and economic benefits. In this paper we ask whether we can identify deprived areas of the city by analysing the flow of people on public transport systems. Cities worldwide are adopting the use of RFID smart-card and sensor based automated fare collection systems to provide access to public transport systems, such as London's Oyster card which accounts for more than 84% of all journeys.[1] With such high coverage, these systems offer a detailed real-time picture of peoples' movement between different parts of the city. To test the hypothesis that automated fare collection systems can be used as rapid, low cost deprivation detectors, we define a number of features derived from transit data and use them to build and test classification models.

BACKGROUND AND RELATED WORK

Since 2000 the UK Office for National Statistics has published, every three or four years, the The Indices of Multiple Deprivation (IMD), a set of indicators which measure deprivation of small census areas in England known as Lower-layer Super Output Areas [13]. These census areas were designed to have a roughly uniform population distribution so that a fine grained comparison of the relative deprivation of different parts of England is possible. The formulation of the IMD follows the principles set out by Townsend [19], in which the author argues that deprivation ought to be defined

*This work was undertaken while at the Computer Laboratory of the University of Cambridge

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

CSCW'13, February 23–27, 2013, San Antonio, Texas, USA.

in such a way that it captures the effects of several different factors. In particular, the IMD consists of the following seven components, here listed with examples (not exhaustive) of factors they measure: Income Deprivation - number of people claiming Income Support, Child Tax Credits or Asylum; Employment Deprivation - number of claimants of Jobseeker's Allowance or Incapacity Benefit; Health Deprivation and Disability - including a standard measure of premature death, rate of adults suffering mood and anxiety disorders; Education, Skills and Training Deprivation - education level attainment, proportion of working adults with no qualifications; Barriers to Housing and Services - homelessness, overcrowding, distance to essential services; Crime - rates of different kinds of criminal act; Living Environment Deprivation - housing condition, air quality, rate of road traffic accidents; and finally a composite measure which is a weighted mean of the seven domains. Note also that the IMD provide relative, rather than absolute measures of deprivation. This is less than ideal, since inequality could be reduced by lowering standards of living in less deprived areas, as well as the more desirable reverse, and with a relative measure we would be unable to determine which has occurred.

The indices have been used by central government as criteria for allocating resources to regeneration programmes and grants to community groups; by local authorities to identify areas in need of support or intervention; and also by residents to hold authorities to account [13]. However, not disregarding the clear utility of this kind of information, one shortcoming is the time frame associated with collating and publishing the data. For example, the latest version of the English IMD, which we use in this study, was published in 2010 yet pertains to data mostly from 2008 and even partly from as early as 2001. To overcome this limitation, researchers have recently proposed several techniques for data collection from ubiquitous sources, such as social media, phone call records and smart card sensors, in order to provide low cost, real-time proxy measures for community well being. For example, Kramer [10] found that the difference between the number of positive and negative words used in Facebook status updates covaries with self-reported "satisfaction with life" in the US. Similarly, Quercia *et al.* found that sentiment expressed in tweets [16] and the topic of tweets [17] in London, aggregated by the area associated with the tweet or Twitter profile, correlates with the composite IMD score of that area. A limitation noted in these works, however, is the large demographic bias of Twitter users. The majority of Twitter users are male, under 35 and with a relatively high income. Mislove *et al.* [14] also suggest that the ethnicity of twitter users in the US is not representative of the general population. Similarly, although Facebook has a more even gender distribution, in the UK around 60% of users are less than 35 years of age.[2]

Research that does not suffer such a population bias involving the IMD includes Eagle *et al.* [5], in which the authors derive a measure of communication diversity from phone call records in England, and find that higher diversity (i.e., the more geographically dispersed a person's social connections) tallies with the composite IMD score aggregated to telephone exchange areas. We adapt this diversity measure to the urban mobility case, but we do not aggregate IMD census areas, since the scores of adjacent areas can differ significantly, thus we offer a finer grained analysis. Furthermore, we do not limit our analysis to the composite IMD score, but additionally investigate each of the seven domains separately.

Other work that analyses the relationship between urban public transport use and composite IMD score is Lathia *et al.* [12], in which the authors find that more deprived areas tend to receive passenger flow from a higher number of other areas compared to less deprived areas, and they also uncover some evidence of social segregation. However, the focus of [12] is the investigation of social mixing and homophily based on trip data. Thus far, no work has been done which explicitly attempts to mine automated fare collection data in order to measure and predict urban social deprivation. In the following sections we introduce the data analysis and classification methods we adopt to do just this.

MINING TRANSIT DATA

In this section we first introduce the dataset and its context, before describing how we extract features of passenger flows and the hypotheses guiding the process.

London and Oyster Card Data

The public transport system in London consists of several interconnected subsystems, incorporating multiple modes of transport. These include 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. In 2003, the operators of this system, Transport for London, introduced an RFID-based technology, known as Oyster card, which serves to replace traditional paper-based magnetic stripe tickets. Oyster cards offer access to the entire multimodal system and thus have the potential to provide a complete picture of public transport in London. We are presently limited to data pertaining to the rail subnetworks, which include a total of 588 stations. Our initial dataset 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 a user with anonymised 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 made by 5.2 million users.

Formulating the Hypotheses

Our goal is to successfully identify areas in London with high deprivation, so we first need to find characteristics of passenger flow patterns, aggregated at the station level, which relate to the IMD scores at the station location. Specifically, we derive parameters related to the following hypotheses based on the literature.

[2] http://www.insidefacebook.com/2010/06/08/whos-using-facebook-around-the-world-the-demographics-of-facebooks-top-15-country-markets/ - retrieved 29/05/2012

1. *Passenger flow is affected by deprivation.* Like an fMRI scanner detecting poor brain function when blood-oxygen flows are not as expected, we propose to capture poor mobility 'function' by measuring the residual between observed flow sizes and those estimated by a simple yet widely used interaction model. This kind of model often incorporates additional parameters to account for the influence of socio-economic factors (e.g., [4, 7]), which leads us to surmise that deprivation will likely impose restrictions on human mobility that we can measure.

2. *Modality choice is affected by deprivation.* The London Travel Demand Survey [1] reports that income is strongly related to public transport use. Compared to rail and tube journeys the rate of bus use tends to increase as household income decreases, and conversely, the rate of car use tends to increase with higher household income. Additionally, the survey finds that mode choice is related to age and disability. Thus, we expect there to be a population bias in rail users related to the level of deprivation in their home areas.

3. *Diversity increases well being.* We expect diversity in the places people visit to reflect the diversity of their social ties, since the more diverse one's contacts are, the less likely they are to be concentrated in a small number of places. Eagle *et al.* [5], using phone calls as a proxy for social connections, have shown that individual level social and economic benefits of social network diversity scale to the population level. Here we use individual travel patterns as a proxy for a social network, and we expect to see a relationship between diversity of travel behaviour and IMD scores.

Processing the Data

We begin by constructing an $N \times N$ matrix F such that N is the number stations in the dataset and $F_{i,j}$ is equal to the average daily number of unique users who have made a trip between stations i and j. We do not take into account direction of travel, so $F_{i,j} = F_{j,i}$. We also find M_i, the set of users who reside near to station i. Since users in the dataset are anonymous we do not know where they live, so instead we infer their home stations using a ranking method that exploits the regularity of human travel patterns [6]. For each user u, we rank station o based on the frequency with which user u has departed from o. To distinguish genuine London residents from occasional visitors, we prune trips not departing within the morning peak period, 6:30am to 9:30am, 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, we then compute a ranking vector $R_u = [r_1, \ldots, r_N]$, where r_k is the number of times u has departed from station o_k (with o_1 being the most frequently visited origin station). We then assign users a home station according to the following set of rules applied 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 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, $m_i = |M_i|$ is the number of users who have i as a home station. The above steps discard 76% of users whose travel records do not reveal any preferential origin station (case (d)), but what is left still amounts to more than 1.2 million users.

Feature Extraction

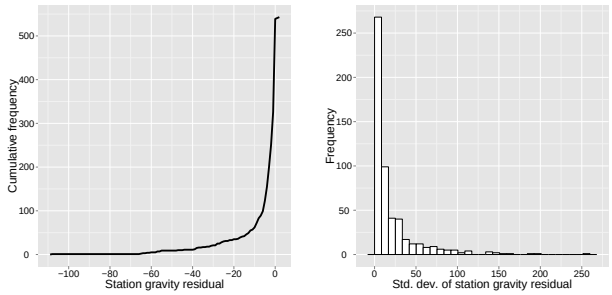
Metrics for hypothesis 1. Next we derive the first set of features which involves using a gravity model to estimate the number of travellers moving between each pair of stations. First introduced by Zipf in 1946 [22], gravity models rest on the hypothesis that the size of flow between two areas is proportional to the mass (i.e., population) of those areas, but decays as the distance between them grows. Despite some criticisms (see for example [18]), the model has been successfully used to describe 'macro scale' interactions (e.g., between cities, and across states), using both road and airline networks [3, 8] and its use has extended to other domains, such as the spreading of infectious diseases [2, 21], cargo ship movements [9], and to model intercity phone calls [11].

Here we posit that a gravity model can be used to estimate passenger flow at the *intra*-city level. The model takes the form:

$$F_{i,j}^{est} = g\frac{m_i m_j}{d_{i,j}^2} \tag{1}$$

where $F_{i,j}^{est}$ is the estimated flow, or number of users moving between stations i and j, g is a scaling constant fitted to the data, and $d_{i,j}$ is the distance between them, for which we use the mean travel time between i and j computed from the transit data. Flows between areas with large mass (large number of home users) and at short distances are predicted to be large, whereas flows at longer distances or between areas with low mass are predicted to be small. Overall, the correlation between the observed traffic flows and gravity model estimates, measured with the Pearson Correlation Coefficient, is as high as .72, which suggests that overall the gravity model provides a good description of the movement of passengers between stations, but also that there is still a significant amount of variation not accounted for by the model.

We posit that this unexplained portion is due to prevailing socioeconomic factors. Thus, we are interested in where the model *fails* to fit well, that is, the cases where the residual (prediction error) between the observed and estimated edge weight is high. For example, the residual between observed and estimated flow between London Bridge and Canada Water, two areas with similar IMD scores, is just 7.9. Conversely, for Liverpool Street and Bethnal Green, which are a similar distance apart but with very different IMD scores, the residual is -1074.742. Figure 1(a) shows the cumula-



(a) Station Gravity Residual (b) Station Gravity Residual Std. Dev.

Figure 1. (a) Cumulative distribution of the mean of gravity residuals at each station. Note that less than 1% of stations have a positive gravity residual, meaning that the gravity model tends to overestimate flows (this is not a plot of G^{μ} which is the mean *absolute* residual); (b) frequency distribution of standard deviations of station gravity residual G^{σ}.

tive distribution of the mean residual at each station. We see that not only is there an acute tendency for the gravity model to overestimate, but also a significant number of stations exhibit a large negative mean residual. As parameters in our classification model, for station i we compute the mean (Equation 2) and standard deviation (Equation 3) of the absolute residual on the edges connected to i:

$$G_i^{\mu} = \frac{1}{k_i} \sum_{j \in S_i} |F_{i,j}^{est} - F_{i,j}| \tag{2}$$

$$G_i^{\sigma} = \sqrt{\sum_{j \in S_i} (F_{i,j}^{est} - F_{i,j})^2} \tag{3}$$

where S_i is the set of stations such that $F_{i,j} > 0$, $\forall j$, and $k_i = |S_i|$. For example, Drayton Green has below average IMD scores and $G^{\mu} = 0.282$ and $G^{\sigma} = 0.809$. In contrast Canary Wharf, which has above average levels of deprivation despite being a busy financial business centre, has $G^{\mu} = 38.295$ and $G^{\sigma} = 117.032$.

Metric for hypothesis 2. The next feature we derive aims to exploit the previously mentioned discrepancy in bus usage between those in high and low income brackets. All else being equal, we would expect m_i to be proportional to the size of the population around i. Thus, we define the population bias as the residual between the observed number of home users, m_i, and that estimated from the population. First we define the population P_i around station i as the sum of a fraction of the population of each census area whose population-weighted centroid is within $1km$ of i. We choose a $1km$ radius based on London Travel Demand Survey finding that the percentage of journeys made by walking is 90% for journeys less than $0.5km$, 60% for 0.5-$1km$ and dropping to below 25% for longer journeys [1]. The size of each fraction is determined with a probability proportional to the distance between the centroid and the station. This method allows the population of each census area to be distributed between several nearby stations which reflects the fact that a person's closest station may not always be on the shortest path to a given destination. Finally, the formula for population bias at i is:

$$Q_i = \log(m_i) - (\alpha + \beta \log(P_i)) \tag{4}$$

where coefficients α and β are fitted by linear regression. Large negative values of Q_i are associated with a bias towards other forms of transport. Large positive values on the other hand are associated with a bias toward using the rail system. Figure 2(a) shows the distribution of population bias, which is roughly normal. We also include the mass, m_i, of the census area in order to control for population size.

Metric for hypothesis 3. Finally, we define two measures of diversity of travel behaviour. For those users for whom we have assigned a home station we follow [5] and measure the diversity of user u's connections as:

$$diversity(u) = \frac{-\sum_{i \in S_u} v_{u,i} \log(v_{u,i})}{\log(|S_u|)} \tag{5}$$

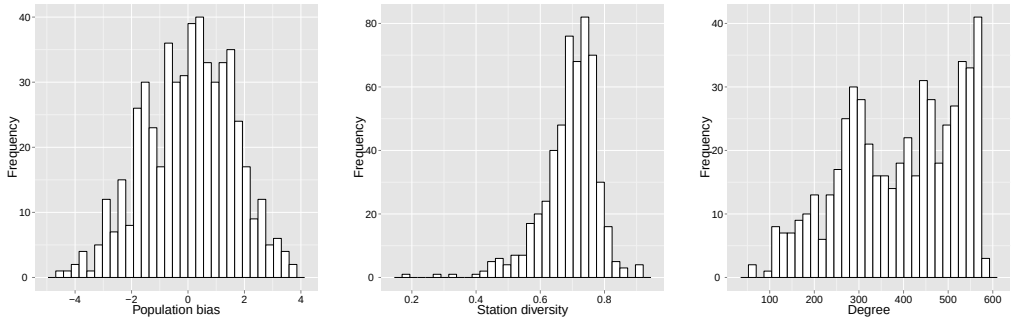
where S_u is the set of stations visited by u, and $v_{u,i}$ is the proportion of all u's visits to station i. The numerator in this equation is the Shannon entropy associated with u's travel behaviour, which is normalised by the log of the number of stations u has visited. Users who visit many places with equal propensity have high diversity, whereas users who tend to concentrate their visits to a few places are considered to have low diversity. The user diversity at station i, denoted H_i, is then the mean of the diversity of all users assigned to i.

$$H_i = \frac{1}{m_i} \sum_{u \in M_i} diversity(u) \tag{6}$$

As a second measure of the diversity of a station, we also include the degree of station i, or the number of other stations to which it is connected, k_i. We include this as we expect the number of connections an area has to increase opportunities for social and economic development, thus lessening the likelihood of deprivation. We include the distribution of station diversity and degree in Figures 2(b) and 2(c) respectively.

TESTING THE HYPOTHESES

To determine whether our hypotheses are valid, we use linear regression analysis. At each station the 8 IMD scores (7 domains and composite) are defined as those of the census area whose population weighted centroid is closest. Figure 4(a) shows the geographical layout of census areas in London. Areas that contain a railway station are coloured according to composite IMD score, where 1 is most deprived. For these areas, each of the 8 IMD scores then becomes the outcome variable for regression, with the features described above in Section (and summarised in Table 1) as the predictor variables. Where necessary, predictor variables undergo a logarithmic transformation, and in addition, we take the z-score of each variable, with allows us to judge the relative influence of each factor in terms of their unit standard deviation. For brevity, in Table 2 we present only the results of the composite IMD score, Living Environment, which has the highest R^2 out of the composite plus 7 domains, and Crime, which we include to demonstrate the variation in the



(a) Population Bias (b) Station Diversity (c) Degree Distribution

Figure 2. Frequency distributions of (a) population bias Q; (b) station diversity H; and (c) station degree k.

F_i	H1	Total flow of station i.
G_i^μ	H1	Mean absolute gravity residual at i.
G_i^σ	H1	Standard deviation of gravity residual at i.
Q_i	H2	Population bias at i.
m_i	H2	Mass of i, or number of home users.
H_i	H3	Mean user diversity at i.
k_i	H3	Degree of station i.

Table 1. List of parameters for the classification models - column 2 gives the corresponding hypothesis.

influence of each feature. In all three examples, G^μ and G^σ, which correspond to hypothesis 1, as well as Q, which corresponds to hypothesis 2, are significant. The sign of the coefficient of G^μ is negative, which indicates that deprivation is expected to increase as this factor decreases. Recall that G^μ is the mean of the absolute value of residuals, and from Figure 1(a) we see that the vast majority of residuals are negative, so larger values of G^μ correspond to overestimates. This means that the more the gravity model overestimates the flows in and out of a station, the more likely that station is in a deprived area. Therefore, this supports our hypothesis that deprivation may represent a restriction to travel, although we cannot infer a causal link. The variance in gravity model error, G^σ, has a positive sign, meaning that deprivation tends to increase as this factor increases. In other words, in cases where the gravity predicts flows between areas to be similar, but in fact they differ, we also see a rise in deprivation - a result which is arguably also in support of hypothesis 1.

The coefficient of transit mode bias, Q, has a negative sign; however, unlike G^μ, this factor has a mean of approximately zero, even before the z-score is taken. So in this case, when more people use the tube than expected (when Q is positive) the regression model expects deprivation to be lower. On the other hand, when less people use the tube than expected (when Q is negative) the model expects deprivation to be higher - this supports hypothesis 2. The support for hypothesis 3, however, is less clear at this stage. Diversity, H, is significant and positive for Composite and Living Environment, which tells us to expect deprivation to increase as diversity increases - the opposite of our hypothesis - although the coefficient of H for living environment is very small. A possible explanation is that diversity of travel behaviour is not a reliable proxy for social network diversity (recall that this was part of the motivation behind hypothesis 3), but instead it may reflect the need of residents to travel to several different places due to lack of provision in their local area. In contrast, we find that the coefficient of the degree of a station, k, is negative, which is in line with our hypothesis that degree may reflect a higher level of economic opportunity (although k is not significant for Living Environment or Crime). In any case, with the exception of Living Environment, the values of R^2 indicate that only around $9 - 13\%$ of the variance in deprivation can be explained by the flow attributes. For this reason, we also tested for interactions between the attributes by including all pairwise products in a linear model, the results of which are shown in Table 3.

For reasons of space, we do not attempt a detailed interpretation of each of the coefficients, and again we show only three example domains. Note, however, that each of the features derived from passenger flow has a significant effect at some point, either alone or in combination. Furthermore, the R^2 values have increased, suggesting that the features explain up to 20% of the variance in deprivation - given the complexity of the problem, we believe this is a good result. An interesting result is that none of the interaction terms are significant for Living Environment, suggesting that either the

	Composite	Liv. Env.	Crime
Intercept	.361 ***	.487 ***	.583 ***
F	.252 *	.070	-.086
G^μ	-.180 ***	-.111 ***	-.060 *
G^σ	.104 *	.231 ***	.109 **
Q	-.141 **	-.229 ***	-.089 **
m	.074	.074	.162 ***
H	.022 *	.001 ***	.009
k	-.116 ***	.031	-.045
R^2	.092 ***	.344 ***	.085 ***
Adjusted R^2	.080 ***	.336 ***	.073 ***

Table 2. Estimated coefficients of Composite, Living Environment and Crime domains (Significance codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

	Composite	Liv. Env.	Crime
Intercept	.369 ***	.467 ***	.583 ***
F	.449 ***	.232 *	-.153
G^{μ}	-.248 ***	-.226 ***	-.020
G^{σ}	.110	.221 ***	.106 *
Q	-.044	-.199 ***	-.085 *
m	-.202 *	-.017	.142 *
H	.035 **	-.003	.018
$F : G^{\mu}$	.216	.102	.303 *
$F : G^{\sigma}$	-.466	-.171	-.624 *
$F : Q$	.705	.205	.776 **
$F : H$	.221 *	.105	.266 **
$G^{\mu} : H$	-.084	-.029	-.104 *
$G^{\mu} : k$	-.343 *	-.071	.012
$G^{\sigma} : m$	.312	.003	.568 *
$Q : m$	-.152 **	-.069	-.124 **
$m : k$	.169	.145	.288 *
$H : k$	-.073 *	-.041	-.112 *
R^2	.199 ***	.392 ***	.195 ***
Adjusted R^2	.154 ***	.358 ***	.151 ***

Table 3. Estimated coefficients of Composite, Living Environment and Crime domains, including pairwise interactions. Predictor variables which no significance in any of the three examples have been omitted. (Significance codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

linear relationships found above exhaust the effect of the features on deprivation in this domain, or, on the contrary, the dependency between each feature is more complex.

Finally, we repeat the above analysis but this time we prune the data points that fall in second and third quartiles of each domain score, thus, we concentrate on the extreme examples and effectively treat the intermediate cases as noise. Note that different stations will appear in the pruned sample for different IMD domains. At this point there are two things worth pointing out: the values of R^2 have risen, which shows that by treating the middle quartiles as noise we are able to discern a stronger dependency between the flow features and deprivation in each of the three examples, and the significance of some features has changed. In particular, H is no longer a significant factor by itself in any of the examples, and for the Living Environment domain we now find that some interaction terms are significant. In general, the change in significance of various attributes suggests that as circumstances change in an area, different factors come in to play, and serves to highlight the complexity of the situation.

In conclusion, the nature of the relationship between the flow features we derived from transit data has thus far evaded a precise description. Nevertheless, we have seen strong support for hypotheses 1 and 2, with a clear dependency between the corresponding features and deprivation in various domains. It is less clear at this stage whether or not hypothesis 3 has any validity.

We next describe how we use the features of passenger flow to build classification models which can identify areas of high deprivation.

	Composite	Liv. Env.	Crime
Intercept	.382 ***	.504 ***	.586 ***
F	.434 *	.103	-.088
G^{μ}	-.351 ***	-.194	-.119 *
G^{σ}	.223 *	.387 ***	.162 *
Q	-.238 **	-.346 ***	-.134 *
m	.122	.057	.252 **
H	.037	.010	.005
k	-.199 ***	.082	-.079 ***
R^2	.160 ***	.508 ***	.146 ***
Adjusted R^2	.138 ***	.494 ***	.123 ***

Table 4. Estimated coefficients of Composite, Living Environment and Crime domains after pruning the second and third quartiles (Significance codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

	Composite	Liv. Env.	Crime
Intercept	.422 ***	.468 ***	.595 ***
F	.728 **	.587 **	-.192
G^{μ}	-.454 ***	-.558 ***	-.054
G^{σ}	.371 **	.399 ***	.186 *
Q	-.053	-.235 ***	-.161 *
m	-.451 **	-.210	.228
k	.116	.113 *	-.029
$F : G^{\sigma}$	-.927	-.662	-1.103 *
$F : m$	-.417	-1.609 *	-.175
$F : k$	-.316	.537 *	-.304
$G^{\mu} : k$	-.457	-.658 **	.027
$G^{\sigma} : m$	.078	.404	.499
$Q : m$	-.373 ***	-.028	-.211 **
$H : k$	-.033	-.023	-.115 *
R^2	.345 ***	.587 ***	.283 ***
Adjusted R^2	.269 ***	.539 ***	.200 ***

Table 5. Estimated coefficients of Composite, Living Environment and Crime domains, including pairwise interactions after pruning the second and third quartiles. Predictor variables which no significance in any of the three examples have been omitted. (Significance codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

IDENTIFYING DEPRIVED AREAS

We formulate the problem as one of binary classification, in which the goal is to identify areas which fall into the top quartile of each of the 8 IMD scores (higher scores corresponded to higher deprivation). We define the response variable as $y_i = 0$ if station i is in the 1st quartile and $y_i = 1$ otherwise. As before, we prune the middle quartiles, since we are interested in classifying the extreme examples. In so doing, we also ensure a roughly $1 : 1$ ratio of positive to negative examples in each class, as opposed to a $1 : 3$ ratio had we retained the 2nd and 3rd quartiles. We compare the results of two linear regression models, the first with each feature as a predictor variable (LR1), and a second model with each pairwise interaction included (LR2). Since the response variable is $y = \{0, 1\}$ we transform the output of the regression models, y', in the following way:

$$y = \begin{cases} 0, & y' \leq 0.5 \\ 1, & y' > 0.5 \end{cases} \tag{7}$$

	Precision	Sensitivity	Specificity	R^2
Composite	.597	.634	.666	.109
Income	.609	.527	.692	.127
Employment	.571	.429	.699	.084
Health	.642	.506	.687	.099
Housing	.714	.605	.747	.184
Crime	.771	.782	.764	.162
Liv. Env.	.819	.771	.826	.521
Education	.702	.753	.649	.141

Table 6. Classification scores for LR1

	Precision	Sensitivity	Specificity	R^2
Composite	.803	.666	.848	.353
Income	.731	.740	.734	.358
Employment	.725	.673	.758	.301
Health	.760	.704	.772	.354
Housing	.742	.712	.799	.428
Crime	.777	.784	.779	.372
Liv. Env.	.852	.832	.834	.621
Education	.687	.787	.643	.312

Table 7. Classification scores for LR2

	Precision	Sensitivity	Specificity
Composite	.805	.733	.810
Income	.705	.729	.733
Employment	.694	.701	.718
Health	.779	.736	.648
Housing	.779	.662	.818
Crime	.761	.756	.769
Liv. Env.	.902	.793	.928
Education	.644	.723	.615

Table 8. Classification scores for SVM

Since our previous analysis points toward a complex relationship between the flow features and deprivation, we also try a support vector machine (SVM) with a radial basis function for the kernel function. The SVM has the advantage that by projecting the input variables onto a higher dimensional space, it is better able to split the input data. The downside to this, however, is that we cannot interpret the role played by each of the predictors, unlike linear regression, for which we can determine the significance of each. For each model, we randomly split the data into a 90% training set and 10% test set, and cross validate by repeating this 10 times.

Performance measures. The performance of each model is assessed using three measures:

$$precision = \frac{TP}{TP + FP}$$

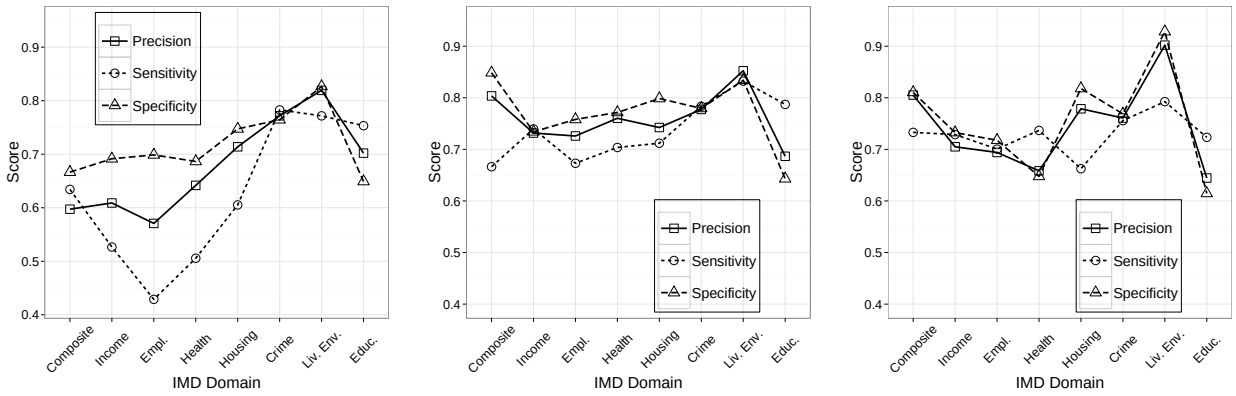
$$sensitivity = \frac{TP}{|P|}$$

$$specificity = \frac{TN}{|N|}$$

where TP is the number of true positives, or correctly identified high deprivation stations; FP is the number of false positives, or stations incorrectly identified as having high deprivation; and $|P|$ and $|N|$ are the total number of positive and negative cases respectively. In the present context, precision measures the proportion of areas classified as highly deprived which are in fact highly deprived. If we assume that a verification cost would be incurred for each area classified as highly deprived, precision would represent the confidence that this cost would not be wasted. It is therefore important that this score is high. Sensitivity (or recall rate) is the proportion of positive examples (stations with high deprivation) in the test set that are correctly identified (i.e., a sensitivity of 1 means all high deprivation areas have been identified). This is perhaps the most important performance measure in the present context, since the idea is to quickly identify areas in need of intervention. Low sensitivity would mean several areas go unreported, thus negating the utility of this kind of classifier. On its own, however, sensitivity is unreliable, since a classifier which labels all examples as positive will have a sensitivity of 1. Specificity is the proportion of negative examples correctly identified as such. High specificity means that for an area identified as deprived, there is a high probability that it is deprived. For the linear regression models we also give the mean value of R^2 for the training data, which represents the proportion of variance in the response variable explained by the predictor variables. This gives an idea of how well the model fits the training data.

Results. We report the mean of the 10 iterations for LR1, LR2 and SVM in Tables 6, 7 and 8 respectively, and for easier visual comparison we have plotted precision, sensitivity and specificity for each model in Figures 3(a), 3(b) and 3(c). The performance of each model can be interpreted relative to a baseline random classifier, which after a sufficient number of iterations averages out with a precision, sensitivity and specificity of 0.5. As expected, the most easily predictable form of deprivation is the Living Environment component. This is the only domain in which LR1 performs well, with high precision, sensitivity and specificity. Note that R^2 differs from that reported in Table 2 since it is computed against the response variable $y = \{0, 1\}$, as opposed to $y = (0, 1)$.

Overall, the best results come from LR2 and SVM. For the composite IMD score, SVM correctly identifies high deprivation 80.5% of the time, and correctly identifies low deprivation 81% of the time. Figures 4(b) and 4(c) provide a geographic illustration of the performance an example SVM model for the composite IMD domain. Each census area is coloured according to class (i.e., bottom quartile or top quartile). Note that only census areas containing stations, and which fall in the 1st or 4th quartile are coloured, hence the map is very sparse. Interestingly, LR2 achieves higher specificity in the composite score, as well as higher sensitivity than SVM in Living Environment, Housing, Crime and Education, suggesting that it may be pertinent to test a mixed classification model in order to achieve the best possible results. A pleasing result is that scores are fairly high across all domains, which demonstrates the potential for identifying specific kinds of problems, rather than just an overall indication of community well being.



(a) LR1 Scores (b) LR2 Scores (c) SVM Scores

Figure 3. Graphical representation of classification scores.

	Precision	Sensitivity	Specificity
Hypo. 1	.640	.539	.702
Hypo. 2	.716	.413	.816
Hypo. 3	.679	.662	.643
Hypo. 1 + 3	.774	.548	.856

Table 9. Performance of SVM classification using features from each hypothesis.

To assess the level of support for our three hypotheses we also trained an SVM model to classify areas in terms of composite IMD, using only features corresponding to each hypothesis separately. Again, we report the average of 10 runs, and the results are presented in Table 9. Each of the performance scores in Table 9 is higher than that expected from a random classifier, with the exception of sensitivity for Hypothesis 2. This tells us that the features related to this hypothesis, population bias Q_i, and mass m_i, struggle to pick out deprived areas and instead tend to classify the majority of areas as undeprived. We can, however, still conclude that Q_i and m_i have predictive power, since an SVM model trained without these features (last row in Table 9) performs less well than one with these features. In general the results suggest that each group of features offers some predictive value, that is, each of the hypotheses has some support, although more work is needed to determine the precise relationship between each of the features and the level of deprivation in a census area.

DISCUSSION

In this section we begin by discussing the main contributions and implications of our work, before examining its limitations and detailing proposals to extend the research.

Implications

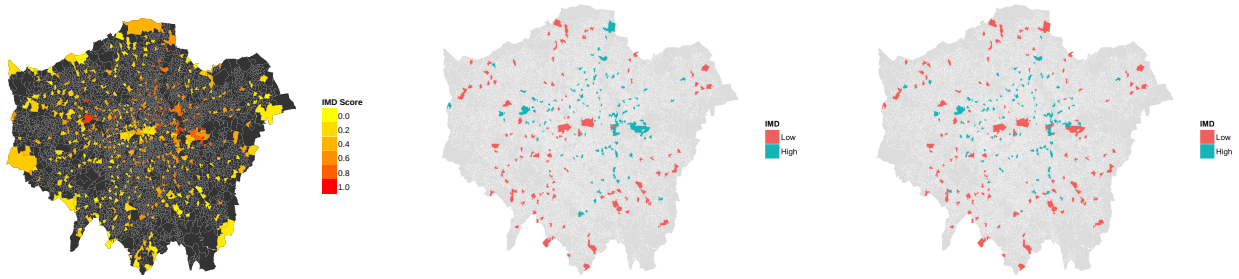
This work adds to the growing literature exploring the ways in which ubiquitous technologies can be used to unobtrusively track the well being of communities [5, 10, 12, 16, 17]. We demonstrated a significant link between census area measurements of deprivation in multiple domains, and patterns of passenger flow in public transport systems. Moreover we have shown that the relationship between passenger flow and urban deprivation is strong enough to build a classification model that uses features extracted from flow data to identify areas of high deprivation. In the spirit of 'smart-cities', predictions derived from transit data could form an element of a 'city dashboard' style application[3], providing real-time information to city planners, policymakers and community members. By providing an early warning, such a tool would dramatically reduce the time frame within which local authorities identify areas of high deprivation, thereby increasing the efficiency with which limited resources are allocated to regeneration and renewal initiatives.

The ability to identify well being and inequality could also benefit communities by enabling residents to asses the effects of regeneration projects and hold local authorities to account. Presently, this is only possible after a number of years, when census based data has been collated and published. Additionally, a city health monitor could help facilitate community-directed campaigns and projects, such as participatory mapping initiatives like `mappingforchange.org.uk`, where residents are asked to vote for what they need the most, and then plan how to spend money accordingly, and `digitaldemocracy.org`, a platform which allows anyone to initiate and garner support for local campaigns. This kind of tool could provide a source of evidence for these types of projects and enable a comparison between areas by assessing the benefits of implementing different policies.

Limitations and Extensions

As previously mentioned, we acknowledge that as we pruned intermediate data points from the training and testing process, there is further work needed to be done in order to verify the results. This would involve classifying all data points in the first three quartiles as not deprived, or introducing a third class to represent moderately deprived areas. Possible ways to improve the models to overcome this limitation in-

[3]e.g., `http://citydashboard.org/london/`



(a) Real composite IMD score (b) True class (c) Predicted Class

Figure 4. a) Census areas containing stations coloured according to real composite IMD score; b) stations which fall in the 1st or 4th quartile for composite IMD, classified as high or low; and c) predicted classifications for the same areas.

clude exploiting other variables available in the Oyster data, such as ticket price and card type (e.g., standard, student, elderly and disabled). We have also begun to develop methods to capture implicit semantic information such as trip purpose, and derive new features such as a gravity residual for each trip type, which we believe are important factors to consider. For example, the rate of commuting may be related to the Employment domain, the number of leisure trips may be related to Income or Crime domains. A further modification would be to measure flows bidirectionally, rather than unidirectionally. This would give us information regarding stations acting as both origin and destination, and allow us to define an asymmetrical gravity model.

A further limitation is that our dataset includes only rail trips and as such we do not have a complete picture of passenger flow patterns in London. Relatedly, the number of areas for which we can make predictions is limited by the geographical coverage of the rail system. Out of a total of 4766 census areas in London, 588 (that is, 11.2%) contain tube or rail stations. With a similar analysis of passenger flow between bus stops we would be able to extend the coverage to the whole of London. Unfortunately we are unable to perform the same analysis on bus trips, since there is no requirement to scan an Oyster card at the end of a bus journey, thus only the origin is recorded. We have, however, developed a proxy for bus usage bias (Q_i) which could potentially be extended beyond the rail/tube network.

We have thus far only tested the models laterally, that is, training and testing on different stations from the same period in time. To establish the genuine worth of building such predictive models, we plan to perform a longitudinal study in which models are tested on newly available data pertaining to trips made several months after the data on which they were trained. Furthermore, a second important verification step is to test our method on data pertaining to different cities, although at present we do not have transit data from other cities at our disposal. Finally, we plan to investigate the utility of supplementing transit data with sources such as Foursquare (and other similar location oriented applications) check-in data, which has been used previously to study global and city scale human mobility patterns [15].

By combining different datasets we can build a multiplex network (i.e. network with multiple types of edge), which may offer additional insights into the relationship between mobility, social structure and well being.

REFERENCES

1. London travel demand survey. Technical Report 1, Transport for London, Greater London Authority, 2009.

2. D. Balcan, V. Colizza, B. Gonalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *PNAS*, 106(51):21484–21489, 2009.

3. A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 101(11):3747–52, Mar. 2004.

4. P. Congdon. Modelling Migration Flows between Areas: An Analysis for London Using the Census and OPCS Longitudinal Study. *Regional Studies*, 23(2):87–103, 1989.

5. N. Eagle and M. Macy. Network Diversity and Economic Development. *Science*, 1029, 2010.

6. M. C. González, C. a. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–82, June 2008.

7. T. Grosche, F. Rothlauf, and A. Heinzl. Gravity models for airline passenger volume estimation. *Journal of Air Transport Management*, 13(4):175–183, July 2007.

8. W. Jung and F. Wang. Gravity model in the Korean highway. *EPL (Europhysics Letters)*, 81, 2008.

9. P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius. The complex network of global cargo ship movements. *Journal of the Royal Society, Interface / the Royal Society*, 7(48):1093–103, July 2010.

10. A. D. I. Kramer. An Unobtrusive Behavioral Model of Gross National Happiness. In *Proceedings of the 28th ACM CHI*, 287–290, 2010.

11. G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, May 2009.

12. N. Lathia, D. Quercia, and J. Crowcroft. The Hidden Image of the City : Sensing Community Well-Being from Urban Mobility. In *Proc. of Pervasive*, 1–8, 2012.

13. D. Mclennan, H. Barnes, M. Noble, J. Davies, and E. Garratt. The English Indices of Deprivation 2010. Technical report, 2011.

14. A. Mislove, S. Lehmann, Y. Ahn, and J. Onnela. Understanding the Demographics of Twitter Users. *Fifth International AAAI*, 554–557, 2011.

15. A. Noulas, S. Scellato, and C. Mascolo. An empirical study of geographic user activity patterns in foursquare. *ICWSM'11*, 70–573, 2011.

16. D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. Tracking Gross Community Happiness from Tweets. In *Proceedings of ACM CSCW 2012*, 2012.

17. D. Quercia, D. O. Seaghdha, and J. Crowcroft. Talk of the City : Our Tweets , Our Community Happiness. In *Proc.of AAAI ICWSM*, 2012.

18. F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 8–12, Feb. 2012.

19. P. Townsend. Deprivation. *Journal of Social Policy*, 16:125–146, 1987.

20. United Nations Human Settlement Program. *State of the World'd Cities 2008/2009*. EarthScan, 2008.

21. C. Viboud, O. N. Bjø rnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science (New York, N.Y.)*, 312(5772):447–51, Apr. 2006.

22. G. Zipf. The P 1 P 2/D hypothesis: On the intercity movement of persons. *American sociological review*, 11(6):677–686, 1946.