Measuring Urban Deprivation from User Generated Content

Alessandro Venerandi
University College London
Gower Street
London WC1E 6BT, UK
alessandro.venerandi.12@ucl.ac.uk

Giovanni Quattrone
University College London
Gower Street
London WC1E 6BT, UK
g.quattrone@cs.ucl.ac.uk

Licia Capra
University College London
Gower Street
London WC1E 6BT, UK
l.capra@cs.ucl.ac.uk

Daniele Quercia
Yahoo Labs
Barcelona, Spain
dquercia@acm.org

Diego Saez-Trumper
Yahoo Labs
Barcelona, Spain
dsaez-trumper@acm.org

ABSTRACT

Measuring socioeconomic deprivation of cities in an accurate and timely fashion has become a priority for governments around the world, as the massive urbanization process we are witnessing is causing high levels of inequalities which require intervention. Traditionally, deprivation indexes have been derived from census data, which is however very expensive to obtain, and thus acquired only every few years. Alternative computational methods have been proposed in recent years to automatically extract proxies of deprivation at a fine spatio-temporal level of granularity; however, they usually require access to datasets (e.g., call details records) that are not publicly available to governments and agencies. To remedy this, we propose a new method to automatically mine deprivation at a fine level of spatio-temporal granularity that only requires access to freely available user-generated content. More precisely, the method needs access to datasets describing what urban elements are present in the physical environment; examples of such datasets are Foursquare and OpenStreetMap. Using these datasets, we quantitatively describe neighborhoods by means of a metric, called *Offering Advantage*, that reflects which urban elements are distinctive features of each neighborhood. We then use that metric to *(i)* build accurate classifiers of urban deprivation and *(ii)* interpret the outcomes through thematic analysis. We apply the method to three UK urban areas of different scale and elaborate on the results in terms of precision and recall.

Author Keywords

Empirical methods, quantitative analysis, socio-economics, user generated content, Foursquare, OpenStreetMap

CSCW'15, March 14 - 18 2015, Vancouver, BC, Canada.

http://dx.doi.org/10.1145/2675133.2675233

ACM Classification Keywords

H.2.8 Database Management: Database Applications—*Spatial Databases and GIS*; H.1.2 Models and Principles: User/Machine Systems—*Human information processing*

INTRODUCTION

The world is undergoing a process of fast urbanization and it is estimated that by 2050 6.2 billion people will live in cities (68% of the total global population) [9]. Although this process is supported by governments as it is expected to bring important advantages (e.g., better and less expansive public services, better living standards due to the concentration of economic activities) [30], recent research has also shown that inequality is dangerously on the rise, with some areas benefiting substantially more than others from public investments and economic growth [28]. Quantifying urban poverty promptly, and at a fine level of spatial granularity, has thus become a priority for governments worldwide, so to be able to monitor the impact of urbanization, and to make data-driven decisions as to how to allocate limited financial resources for regeneration projects.

Traditionally, socioeconomic deprivation has been measured using data acquired through household surveys; while such data is semantically rich, it is also very expensive to obtain and process. As a result, it is acquired with a rather low frequency that varies from few years for developed countries like the UK, to several years for developing countries like Cote d'Ivoire. To remedy this, computational social scientists have started to develop new methods that aim to automatically derive metrics of deprivation from alternative data sources, that afford finer spatio-temporal granularity than survey data. Data sources used to date span from call detail records (e.g., [10, 25, 40]), to satellite images (e.g., [11, 12, 29]), to transit data (e.g., [41]). A common limitation to all these methods is the reliance on datasets that are very difficult to obtain, thus severely limiting the applicability of the methods themselves.

In this paper, we propose a new method that aims at accurately computing urban deprivation at a fine-grained spatio-temporal granularity, while relying on easily accessible datasets. More precisely, in terms of datasets, our method relies on user-generated content that captures which urban features are present in a neighborhood. Examples of such

datasets are Foursquare and OpenStreetMap. We made this choice inspired by qualitative works in the public health domain that have found important relationships between the presence of certain urban elements in a given area, and the socioeconomic well-being of its residents. For example, researchers have found that health-promoting amenities (e.g., golf courses in Australia [14], fitness centers and dance facilities in USA [32]) are more concentrated in well-off areas; on the contrary, potentially health-harmful resources (e.g., fast food outlets in England and Wales [8]) are more concentrated in poorer areas. From these datasets, our method automatically extracts a metric called *Offering Advantage* that intuitively reflects which urban elements are distinctive features of each neighborhood. Using correlation analysis, we prune down these features so to consider only those that are significantly correlated with urban deprivation; we then use those significant features to build classifiers of urban deprivation and finally, by means of thematic analysis, interpret the outcomes. We illustrate how to apply the method in practice in three UK urban areas of different scale (i.e., Greater London, Greater Manchester, West Midlands); finally, for these three case studies, we elaborate on the precision and recall of the results.

In the reminder of the paper, we first provide an overview of related works in this domain. We describe the datasets, and the method developed to leverage them. We then present the results of our evaluation, before concluding the paper with a discussion of implications, limitations and future work.

RELATED WORK

In an attempt to measure socioeconomic deprivation at a fine level of spatio-temporal granularity, researchers have started to develop computational methods that automatically mine a variety of data sources, looking for significant and strong signals of deprivation. To this end, three main sources of data have been used.

Call Detail Records (CDRs). These are records about calls and text messages measured by telecommunication providers for billing purposes. CDRs contain information about time, duration, caller ID, callee ID and location of the antenna tower through which the call or the text message is sent. Features extracted from these datasets have then been used to assess socioeconomic well-being of populations. A pioneering work in this domain has been conducted by Eagle *et al.* [10]; they studied the relationship between CDRs from land lines and mobile phones in England and the Index of Multiple Deprivation (IMD). Their results highlight a strong correlation between call network diversity and deprivation, confirming the hypothesis that having a varied set of contacts is a signal of socioeconomic well-being. A few years later, Mao *et al.* investigated the relationship between features of calls in Cote d'Ivoire and some economic indexes of ten areas of high economic activity [25]. They discovered that the ratio of outgoing calls per area, relative to the total of outgoing plus incoming calls, has high correlation with annual income. Smith *et al.* analyzed the same mobile phone dataset and found that features of network diversity and introversion (i.e., ratio of within-area calls vs. inter-area calls) strongly correlate with deprivation [40]. A common limitation to these studies is the difficulty to gain wide access to CDR data, as telecommunication operators do not tend to make that data publicly available.

Satellite imagery. A completely different approach has looked into analyzing patterns from satellite images, in order to map economic development. In particular, researchers have extracted a feature called Night Time Light (NTL) from images, that is, the total surface lit during night time. Elvidge *et al.* found a correlation between NTL and countries' Gross Domestic Product [11, 12]. Similarly, Noor *et al.* studied the relationship between NTL and a composite index of wealth for several administrative regions of some African countries [29]. The correlations found were initially high; however, more recent research showed far lower correlations. These findings suggest that, as the penetration of electrical lightning reaches saturation, the signal present in these datasets disappears. Satellite images are as hard to get as CDRs; furthermore, the methods seem only applicable in under-developed countries up to the point where electricity becomes a commodity.

Transit data. Another source of information which researchers have been analyzing to get insights into deprivation is urban transit data. Smith *et al.* [41] used Oyster Card data (i.e., electronic ticketing system capturing journeys made within the London public transport network) to derive features of mobility flow between areas, and of transport modality choice. They then used these features to build a classification model to identify highly deprived areas, as measured by the UK Index of Multiple Deprivation. Although the model achieves high prediction accuracy, it can only estimate deprivation for 10% of London (i.e., where a tube station is present). Furthermore, similar to works on CDRs data, this line of work requires access to datasets that are very difficult to get hold of (if available at all).

All lines of research above require access to datasets that are very difficult to get. An alternative approach is to rely on more easily accessible user-generated content (UGC) for the same purpose. User-generated content comes in a wide variety of forms, thanks to the big popularity and uptake of social media applications, especially in developed countries. These data sources have been used by researchers to develop a better understanding of our cities. For example, geo-coded tweets have been analyzed to quantify sentiment/mood, as it varies between neighborhoods of different socioeconomic standing [35]; Twitter has also been analyzed to investigate the relationship between the topic of discussion in a certain area, and the deprivation score of that area [38]; Foursquare check-ins have been used to redefine neighborhood boundaries, by categorizing different areas of a city through a clustering model that leverages similarity of urban functions, rather than using super-imposed administrative boundaries [7].

In line with this last stream of research, we propose to use user-generated content to mine urban deprivation. Inspired by previous qualitative works in the public health domain that found correlations between the presence of certain urban venues and deprivation [14, 32], we propose to automatically

extract such urban features directly from easily accessible user-generated content datasets, specifically, Foursquare and OpenStreetMap. In the next section, we describe the datasets, before providing the details of our method.

DATASETS

To conduct this work, we needed access to two types of datasets: on one hand, indicators of socioeconomic deprivation at a fine-grained spatial granularity; on the other hand, detailed records of which physical elements are present in the built environment. We use the Index of Multiple Deprivation for the former, and both Foursquare and OpenStreetMap for the latter.

Index of Multiple Deprivation (IMD)

As measure of deprivation, we use the UK Index of Multiple Deprivation (IMD),[1] computed at the level of small census areas known as Lower-layer Super Output Areas (LSOAs). LSOAs were defined to roughly include always the same number of inhabitants (around 1.500) [27]. IMD is a composite score, calculated as the weighted means of seven distinct domains: income deprivation, employment deprivation, health deprivation, education deprivation, barrier to housing and services, crime, and living environment deprivation. The higher the IMD score, the more deprived the neighborhood, and viceversa; overall, IMD scores follows a normal distribution [27]. For the purpose of our study, we collected IMD scores for three UK urban areas that differ in terms of population size and geographic span, so to test the applicability of our approach to case studies of different scales. Those are: Greater London, Greater Manchester, and West Midlands.

We chose Greater London as an example of large metropolitan city. Greater Manchester and West Midlands are both examples of mid-size cities instead, albeit with a rather different population density. General information about these cities is provided in Table 1.[2]

Urban area	*Population*	*Density*	*Area*
Greater London	8,204,100	5,218 ppl per km^2	1,572 km^2
Greater Manchester	2,685,400	2,105 ppl per km^2	1,276 km^2
West Midlands	2,738,100	3,039 ppl per km^2	902 km^2

Table 1. Population, population density, and area for the three UK urban areas.

Foursquare

Foursquare is a mobile social-networking application launched in 2009 and is also one of the most popular location-based social-networking websites.[3] In Foursquare, when registered users visit a venue, they can 'check-in' on the mobile application to share their location with their friends. In April 2012, Foursquare reported 20 million registered users, with more than 2 billion check-ins [20]. Aside from checking-in at existing venues, Foursquare users can also create new ones. Possible conflicts in the definition of venues are solved in a bottom-up fashion: the more accurate a description, the more likely users will be able to recognize (check-in to) it. Foursquare then attempts to merge multiple descriptions that are likely to refer to the same venue using a Venue Harmonization procedure,[4] which includes the use of developer-contributed geographic databases. At the minimum level, each venue needs to be defined by a pair of latitude/longitude coordinates, a name, and a category (e.g., Church, School, Pub). Janne Lindqvist *et al.* have recently studied why people check-in and found five factors, one of which is particularly relevant to the creation of places [22]: individuals tend to use Foursquare to see where they have been in the past and ultimately curate their own location history. In cities where Foursquare has high penetration, the venues recorded in this service should thus collectively form a well-curated land use dataset. For the purpose of this work, we use the official Foursquare API to crawl all Foursquare venues for the three UK urban areas under consideration.[5] We performed this step between 04/03/2014 and 08/04/2014; a summary of the dataset obtained is reported in Table 2. Given that the three cities show different orders of magnitude in the number of venues, applying our method to them is likely to translate into interesting insights about our method's applicability to different urban contexts.

Urban area	*# Venues*	*# Check-ins*	*# Categories*
Greater London	178.756	26.344.132	503
Greater Manchester	43.874	3.235.174	421
West Midlands	37.370	2.424.546	435

Table 2. Number of Foursquare venues, number of check-ins, and number of Foursquare categories across the three UK urban areas.

OpenStreetMap

OpenStreetMap (OSM) is perhaps one of the most successful examples of geographic crowd-sourcing, with currently over 1.6M users, collectively building a free, openly accessible, editable map of the world.[6] OSM data covers three types of spatial objects: *nodes*, *ways*, and *relations*. Nodes broadly refer to Points of Interests (POIs), ways are representative of roads, and relations are used to group together other objects (e.g., administrative boundaries, bus routes). For the purpose of this study, only nodes are relevant. Similar to Foursquare venues, an OSM node consists of three main attributes: a geographical position (latitude and longitude), a name, and a category (called amenity type in OSM jargon). Differently from Foursquare, OSM categories are not chosen through a given taxonomy, and contributors are free to use whatever words they find most suitable. We downloaded OSM node data for the three cities under consideration on 07/05/2014; summary statistics are provided in Table 3.

Urban area	*# Nodes*	*# Categories*
Greater London	79.343	896
Greater Manchester	24.321	381
West Midlands	27.885	465

Table 3. Number of OSM nodes and number of OSM categories across the three UK urban areas.

Preliminary analysis on Foursquare and OSM datasets for Greater London (Figure 1) shows that categories follow a

[1] `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6871/1871208.pdf`

[2] `http://www.ons.gov.uk/ons/dcp171778_270487.pdf`

[3] `https://foursquare.com/about`

[4] `https://developer.foursquare.com/overview/mapping`

[5] `https://api.foursquare.com`

[6] `http://download.geofabrik.de/`

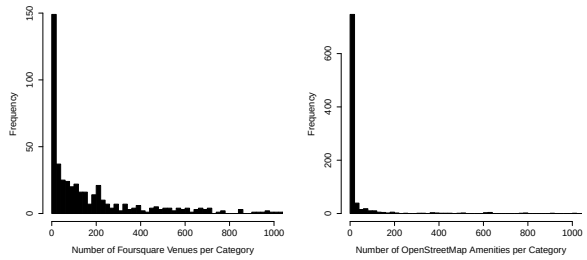

Figure 1. Frequency distribution of Foursquare categories (left) and OSM categories (right).

long-tailed distribution, with few categories having large number of venues/nodes, and many categories with very few venues/nodes instead. Furthermore, we notice that, although Foursquare and OSM conceptually provide the same kind of information (e.g., physical elements present in the urban environment), their users map different things. In fact, while the most frequent Foursquare categories are restaurants, pubs and cafes, the most frequent OSM categories are bus stops, crossings and post boxes. This is not surprising, given the fundamentally different purposes behind these services: Foursquare is mainly used to share locations with which users like to be associated; conversely, OSM is mainly used to capture in full what is present in the world. Our work takes advantage from the complimentary nature of these two UGC datasets, as we hypothesize that more accurate area profilers can be built by combining the peculiarities of the two datasets, rather than using them in isolation, thus potentially unveiling a broader set of features that correlate with urban deprivation.

METHOD

In this section, we describe the method we have developed to estimate urban deprivation from UGC data (Foursquare and OSM in particular). We begin with a description of the adopted spatial unit of analysis, and of the motivation behind this choice. We then define the metric implemented to automatically extract, from UGC data, the urban features characterizing these spatial units. Finally, we provide a step-by-step description of how to apply our method in practice: from correlation analysis to identify urban features related to deprivation, to classifiers to estimate deprivation levels, to thematic analysis to interpret results.

Unit of Analysis

The goal of this work is to measure the socioeconomic deprivation of different areas within a city by capturing the urban elements that are physically present in each neighborhood. To do so, we need to define a spatial unit of analysis that is representative of a city neighborhood. As mentioned in the previous section, IMD is available at a fine-grained spatial granularity, that of LSOAs. However, these units, which have been introduced relatively recently (in 2001), are too small and too arbitrarily defined to be meaningful to urban residents. Indeed, we calculated the average LSOA for Greater London to be no bigger than 33 hectares; furthermore, they are identified through alphanumeric codes (e.g., E01008881) which do not represent neighborhoods as recognized by citizens. We therefore choose a different spatial unit called *ward*. Wards have a much longer history compared to LSOAs (they have existed since the Middle Ages). Wards are defined by the UK Government[7] and represent both electoral subdivisions and ceremonial entities; their geographic extension exceeds that of the LSOAs with an average area size for Greater London of about 250 hectares. Finally, wards are identified through human-understandable toponyms (e.g., Highgate). For all these reasons, we argue that wards, more than LSOAs, are good representations of citizens' neighborhoods. Using official geographic definitions of wards in the UK,[8] we computed 625 wards for Greater London, 215 for Greater Manchester and 163 for West Midlands. For simplicity, in the next sections we will refer to wards as 'neighborhoods'.

Offering Advantage Metric

Having defined our spatial unit of analysis, we now need to profile each ward in terms of the urban features that characterize it. In doing so, we aim to capture not just what urban elements are physically present in a neighborhood, but more importantly what elements make it distinct with respect to other neighborhoods. As shown in the previous section, some categories are much more frequent than others, so that a simple count of what amenities are present is not sufficient to elicit distinctiveness. Rather, we propose to use a metric called *Offering Advantage*, which weights categories by their popularity; intuitively, the presence of one element from an unpopular category is much more significant in profiling a neighborhood than the presence of one element from a very popular category. In practice, this measure relies on a concept used in economics called Revealed Comparative Advantage (RCA) [17]. This is used to measure whether a country exports more of good i (as a share of its total exports), than the average country; if so, then $RCA > 1$.

The RCA of a country c is usually evaluated with this formulation:

$$RCA_{c,i} = \frac{goods_{c,i}}{goods_c} \cdot \frac{world}{world_i}$$

where $goods_{c,i}$ denotes how many goods i are exported by country c; $goods_c$ denotes the total number of goods exported by the country c; $world$ is the total number of goods exported all around the world; finally, $world_i$ indicates how many goods i are exported all around the world.

In our context, this measure reflects to what extent a neighborhood n_k provides more (Foursquare/OSM) POIs of a certain category c_i than the average neighborhood. More specifically:

$$OA(c_i, n_k) = \frac{count(c_i, n_k)}{\sum_{j=1}^{N} count(c_j, n_k)} \cdot \frac{\sum_{j=1}^{N} count(c_j)}{count(c_i)}$$

[7] **http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/electoral-wards-divisions/index.html**

[8] **https://geoportal.statistics.gov.uk/Docs/Boundaries/Wards_(E+W)_2011_Boundaries_(Full_Extent).zip**

where $OA(c_i, n_k)$ denotes the *Offering Advantage* of a (Foursquare/OSM) POI category c_i in the neighborhood n_k; $count(c_i, n_k)$ counts how may POIs of category c_i are present in the neighborhood n_k; N is the total number of (Foursquare/OSM) POI categories; finally, $count(c_i)$ counts how many (Foursquare/OSM) POIs of category c_i are present in the whole urban area. This metric has been recently used in a preliminary study on a Foursquare dataset for Greater London [37] and showed to provide better association with deprivation compared to a raw count of number of POIs for each category in a ward. We thus rely on the same metric to study deprivation and we apply it on two datasets of user generated content and on three cities.

Approach

Having defined our spatial unit of analysis (i.e., UK wards) and the metric we use to quantitatively profile these areas (i.e., *Offering Advantage*), we can now describe our proposed approach.

Correlation Analysis to Identify Significant Urban Features

Each area (ward) is described by a vector that reports the OA metric for each Foursquare/OSM POI category in that area. Although there are several hundred POI categories, not all of them will bear significant signals of deprivation. For example, the very same POI category (e.g., school) may be equally present in well-off and deprived wards. To filter out all those categories that do not consistently signal deprivation within the urban area under consideration, we use correlation analysis between the OA metric automatically derived from user-generated content, and the Index of Multiple Deprivation (IMD) that acts as ground truth in this context. To do so, the following three steps had to be performed: first, we had to reconcile the spatial unit of analysis at which IMD is available (i.e., LSOA), with the spatial unit of analysis used in this work (i.e., ward). We did so by computing the IMD score of a ward as the average of the IMD scores of the LSOAs it spatially contains. One may wonder if, by doing so, we cause significant data loss and inaccuracies. We found this not to be the case, since deprivation scores for LSOAs which belong to the same ward are very consistent (the standard deviation of IMD values related to LSOAs contained within wards is smaller than the corresponding average value, for all wards).

Second, since we are dealing with geographical data, we had to address the problem of spatial auto-correlation in our data. This, in fact, can lead to incorrect conclusions. Spatial auto-correlation is the tendency for measurements located close to each other to be correlated, a property that generally holds for variables observed across geographic spaces [21]. In broader terms, this is the direct quantitative demonstration of Tobler's First Law of Geography, which states that 'everything is related to everything else, but near things are more related than distant things' [43]. When high spatial auto-correlation occurs, traditional metrics of correlation (e.g., Pearson and Spearman) that require independence in the observations cannot be applied. We tested our data for spatial auto-correlation and we found it to be indeed high. To overcome this problem, we used a method familiar to natural scientists [15] and introduced by Clifford *et al.* [6]. This approach addresses the 'redundant, or duplicated, information contained in georeferenced data' [16] – the effect of spatial auto-correlation – through the calculation of a reduced effective sample size. The significance of the correlation coefficients presented in the next section is obtained through the implementation of Clifford *et al.*'s method, partly accounting for spatial auto-correlations.

Third and lastly, since we are performing simultaneous correlation tests with multiple variables (i.e., our POI categories), the chance of incorrectly rejecting the null hypothesis for some of these variables (and thus of obtaining false positive results), increases. To quantify this threat, we implement a statistical control technique commonly used among researchers dealing with datasets comprising a large number of distinct variables (e.g., in genomics) called False Discovery Rate (FDR) [42]. In practice, the method analyses the distribution of p-values of the tested variables, and produces a list of so-called q-values, each varying between 0 and 1, indicating the expected proportion of false discoveries within the list of findings. In the next section, we will report q-values, computed using the FDR method, over the list of variables found to be significantly correlated when using the Clifford *et al.* correlation method.

Note that there exists a temporal discrepancy between the IMD dataset (2011), and the Foursquare/OSM datasets (2014). We expect this discrepancy to have limited impact on the findings we will report in the next section, as IMD values did not significantly vary in the last two reporting periods (i.e., 2008 and 2011).

Classification Tool of Deprivation

Once we have identified the POI categories that correlate with deprivation in the urban areas, we can then build a classifier of deprivation. One might wonder what the advantage of such a classifier is, if we still need IMD scores to be able to identify what urban features signal deprivation in the first place. We envision two uses of this approach that would make it more affordable than running citywide household surveys every (few) year(s): a first approach would require the completion of household surveys only in a small subset (e.g., 25%) of the city neighborhoods; our method would then apply correlations analysis to derive features upon which to build classifiers for the remaining (e.g., 75%) areas. A second approach would see the completion of citywide household surveys, from which to compute IMD scores manually, only once every several years: at these times, correlation analysis would be conducted. Once this is done, for the several years in between surveys, the classifiers would then be used to estimate deprivation at no extra cost. This second approach is based on the assumption that, although the deprivation score of an area may vary year by year (e.g., as a result of processes of urbanization, migration, and gentrification), the relationship between deprivation scores and urban features is much more stable; that is, certain POI types remain concentrated where wealth – or deprivation – is higher (although we do not know whether that is because people move towards area with certain POI types, or because areas with some POI types attract people of certain economic status, or both).

Thematic Analysis to Derive Significant Themes

The correlation analysis potentially identifies tens of Foursquare/OSM POI categories that are associated with deprivation; some of these might be redundant (e.g., bus, bus stop), and some others might be due to chance. The outcome of the classification tool built atop of these categories might thus be fragmented and difficult to interpret. To avoid this, our proposed method requires the application of inductive thematic analysis [4] to the filtered Foursquare/OSM POI categories, so to group together those categories that are semantically related. The result is a very small set of meta features, or themes, that represent simple and distinctive urban characters.

In the next section, we apply our method to three UK urban areas of different scale (i.e., Greater London, Greater Manchester, West Midlands), and discuss precision and recall of the results.

RESULTS

We first illustrate the results of the correlation analysis that is carried out to identify which POI categories (out of the hundreds present in Foursquare and OSM) are significantly associated with deprivation. Second, we build classification tools based on the identified categories, and assess their accuracy compared to a baseline classifier and to a state-of-the-art approach. Finally we illustrate which meta-features emerged from the thematic analysis.

Finding Signals of Deprivation

We explore to what extent we can exploit land use data extracted from user-generated content to get useful insights about the socioeconomic status of city neighborhoods. To this end, we calculated *Offering Advantage* for all Foursquare and OSM categories for each urban area and, through the Spearman's rank correlation coefficient r_s, we correlated it with IMD. Tables 4 and 5 show the number of Foursquare and OSM categories correlated (positively or negatively) with IMD, grouped by strength of correlation. From the analysis of these two tables, we highlight two important observations: first, there is a higher number of categories significantly correlated with deprivation in Foursquare than in OSM; this suggests that most OSM categories are conceptually less associated with socioeconomic aspects of cities. Second, the number of weak-to-moderately correlated categories ($r_s \in [0.2, 0.4)$ or ($r_s \in [0.4, 0.6)$) is high across all cities (15 categories for Greater London, 24 for Greater Manchester, and 34 for West Midlands), suggesting there is indeed a wealth of urban features we can mine from UGC to study deprivation for cities of different scales (in terms of area size, population, and user-generated content). Tables 6 and 7 show the top three categories most positively and most negatively correlated with IMD, using Foursquare and OSM data respectively. Note that, while Foursquare categories provide detailed information about services and facilities (e.g., Student Center, Caribbean), OSM categories tend to give more information about road system elements (e.g., traffic signals, crossing). The two datasets thus offer complimentary information at times, and they should thus be studied together.

	$\lvert r_s \rvert \in [0.05, 0.2)$	$\lvert r_s \rvert \in [0.2, 0.4)$	$\lvert r_s \rvert \in [0.4, 0.6)$
Greater London	23	15	0
Greater Manchester	30	24	0
West Midlands	17	33	1

Table 4. Number of Foursquare POI categories (positively or negatively) correlated with IMD (all results shown are statistically significant, $p <$ 0.05)

	$\lvert r_s \rvert \in [0.05, 0.2)$	$\lvert r_s \rvert \in [0.2, 0.4)$	$\lvert r_s \rvert \in [0.4, 0.6)$
Greater London	4	2	0
Greater Manchester	1	4	0
West Midlands	1	9	0

Table 5. Number of OSM POI categories (positively or negatively) correlated with IMD (all results shown are statistically significant, $p <$ 0.05)

To quantify the expected proportion of false discoveries among the previously identified POI categories, we then followed our proposed methodology and applied the FDR control technique. More precisely, we first ranked our variables by their p-values (from the lowest to the highest), and then computed q-values on the ranked list. Table 8 shows summary results: for each dataset (Foursquare and OSM), and for each city under study (Greater London, Greater Manchester and West Midlands), we report the computed q-values for each quartile of the (ranked) variables. Let us consider Foursquare-related results first: the expected proportion of false positive correlations is less than 9% across all POI categories considered significant for Greater Manchester, and at most 20% for West Midlands. For Greater London, the expected proportion is at most 14% for half of the discovered variables (those with lowest p), though it increases to 40% when looking at the whole set. Results for OSM-related variables are less promising instead: with the exception of West Midlands, where the q values are low over the entire set of variables, for Greater London up to 63% of the findings are expected to be false discoveries, and up to 41% for Greater Manchester. For the purpose of the present study, the Foursquare dataset would thus appear more suited; indeed, the majority of significant POI categories, and the themes derived from them (which we are going to present next), come from Foursquare.

Building Classifiers of Deprivation

We now test whether we can exploit land use data extracted from user generated content to build accurate classifiers of urban deprivation. As we pointed out in the Method section, there are two possible ways to carry out this task: conducting household surveys on a small subset of city neighborhoods and estimating deprivation for the remaining ones; or conducting citywide surveys in one year and estimate deprivation for the subsequent ones. We evaluate the former approach next (we cannot test the latter, as we do not have UGC for different time steps). We proceeded as follows: we selected Greater London as case study (similar results were obtained for Greater Manchester and West Midlands), and divided IMD values into ten deciles, adhering to the methodology applied in the official IMD document [27]. We then randomly split our data in 25% train and 75% of test, thus obtaining 156 neighborhoods for the train test and 469 neighborhoods for the test set respectively. We then built a variety of classifiers that take in input the *Offering Advantage* values of the Foursquare/OSM POI categories that show statistically

	Greater London	*Greater Manchester*	*West Midlands*
Top positively correlated	Caribbean (0.37) African (0.32) Fried Chicken (0.31)	Bus Station (0.32) Residential (0.27) Student Centre (0.24)	Car Wash (0.38) Temple (0.34) Desserts (0.32)
Top negatively correlated	Indian (-0.27) Italian (-0.26) Golf Course (-0.24)	Italian (-0.36) Golf Course (-0.28) Gastropub (-0.26)	Golf Course (-0.42) Salon Barbershop (-0.35) Farm (-0.31)

Table 6. Foursquare categories most (positively and negatively) correlated with IMD. In parentheses, the Spearman correlation values (at statistical significance $p < 0.05$) computed with the Clifford *et al.* method are shown.

	Greater London	*Greater Manchester*	*West Midlands*
Top positively correlated	traffic signals (0.29) crossing (0.25) community center (0.18)	traffic signals (0.25) taxi (0.24)	tram stop (0.31) billboard (0.29) artwork (0.29)
Top negatively correlated	parking (-0.14) garden center (-0.10)	post box (-0.26) kindergarten (-0.22) restaurant (-0.18)	parking (-0.30)

Table 7. OSM categories most (positively and negatively) correlated with IMD. In parentheses, the Spearman correlation values (at statistical significance $p < 0.05$) computed with the Clifford *et al.* method are shown.

Dataset	*Urban area (# categories)*	*1st Qu.*	*Median*	*3rd Qu.*	*Max.*
Foursquare	Greater London (35)	0.04	0.14	0.30	0.40
	Greater Manchester (54)	0.02	0.03	0.05	0.09
	West Midlands (50)	0.06	0.10	0.14	0.20
OSM	Greater London (6)	0.32	0.32	0.63	0.63
	Greater Manchester (5)	0.20	0.21	0.31	0.41
	West Midlands (10)	0.07	0.07	0.14	0.14

Table 8. q-values per quartiles of POI categories, computed using the False Discovery Rate technique.

significant (positive or negative) correlations with IMD values in the training set. Next, we present results obtained using a Naive Bayes classifier – the classifier that, among the tested ones (i.e., Decision Tree j48, Logistic Regression), showed the best performance.

Classification accuracy results are shown in Table 9. We note that the highest Precision and Recall values are obtained for classes a and j, which represent the 10% least deprived and the 10% most deprived neighborhoods. This suggests that our method is most suitable to identify and monitor problematic areas, while performing less well for middle cases. Also, at first glance, Precision and Recall are not particularly high in absolute values. However, in almost all cases, the predicted class only differs of a few positions (two or three) from the actual one, as evidenced by the Confusion Matrix reported in Table 10.

Precision	*Recall*	*F-Measure*	*Class*
0.320	0.356	0.337	a: 10% more deprived
0.189	0.227	0.206	b: from 10% to 20% more deprived
0.200	0.137	0.163	c: from 20% to 30% more deprived
0.135	0.102	0.116	d: from 30% to 40% more deprived
0.125	0.067	0.087	e: from 40% to 50% more deprived
0.143	0.080	0.103	f: from 50% to 60% more deprived
0.165	0.357	0.226	g: from 60% to 70% more deprived
0.194	0.255	0.220	h: from 70% to 80% more deprived
0.214	0.115	0.150	i: from 80% to 90% more deprived
0.262	0.364	0.305	j: from 90% to 100% more deprived

Table 9. Classification accuracy of urban deprivation for Greater London. IMD is subdivided in 10 bins.

At this point, one may wonder how our proposed method performs compared to both state-of-the-art approaches (e.g., [10, 25, 40]) and simpler benchmarks derived from UGC datasets.

a	b	c	d	e	f	g	h	i	j	← classified as
16	11	5	3	3	1	2	1	0	3	a: 10% more deprived
10	10	4	4	4	0	7	3	1	1	b: from 10% to 20% more deprived
4	9	7	7	3	4	7	4	3	3	c: from 20% to 30% more deprived
7	4	4	5	3	3	17	2	2	2	d: from 30% to 40% more deprived
2	6	5	2	3	2	11	6	3	5	e: from 40% to 50% more deprived
4	4	6	2	1	4	12	4	4	9	f: from 50% to 60% more deprived
0	5	0	3	2	4	15	6	1	6	g: from 60% to 70% more deprived
2	3	1	5	1	3	11	12	3	6	h: from 70% to 80% more deprived
4	1	3	2	3	5	4	14	6	10	i: from 80% to 90% more deprived
1	0	0	4	1	2	5	10	5	16	j: from 90% to 100% more deprived

Table 10. Confusion Matrix associated with our classification model.

To answer the former, we compare our classification results with those reported in [40], where public transit data was used to estimate deprivation for Greater London, using the same ground truth data (IMD published in 2011) we rely upon. To ease comparison, since the work in [40] divided the IMD distribution in two bins only (i.e., below and above the median value), we re-computed our classification using these two bins separated by the median value as output. To address the latter, we built a simple benchmark that estimates deprivation of a London ward by means of a Naive Bayes classifier that takes in input the number of Foursquare check-ins, the number of Foursquare POIs, and the number of OSM POIs present in that area; the intuition behind this benchmark is that, the higher the number of check-ins and POIs in a ward, the less deprived the ward is.

Table 11 shows results for our model, together with the performance gain over the basic benchmark. Our model reaches a Precision between 0.763 (for above-median deprivation) and 0.713 (for below-median deprivation); the best performing classifier presented in [40] achieved an overall Precision of 0.805; however, note that such result only holds for 10% of the wards in London (where a tube station is present), while our results cover the whole of Greater London. By taking these two observations together, we argue that the performance of our classifier is indeed comparable to state-of-the-art approaches that require access to datasets that are not publicly available. As for the performance of our model compared to the simpler benchmark, we observe significant improvements (shown in parentheses in Table 11) for both Precision and Recall in both classification classes. This result demonstrates the suitability of the *Offering Advantage* metric over simple counts of check-ins and POIs.

Precision	*Recall*	*F-Measure*	*Class*
0.763 (+41%)	0.692 (+17%)	0.726 (+28%)	50% more deprived
0.713 (+37%)	0.780 (+39%)	0.745 (+38%)	50% less deprived

Table 11. Classification accuracy of urban deprivation for Greater London. IMD is subdivided in 2 bins. In parenthesis, the percentage difference w.r.t. the results of a basic benchmark is shown.

Deriving Themes

The previous results show that our method has competitive classification accuracy. To interpret those results, we conduct a thematic analysis [4] of the POI categories identified through correlation analysis, and group them together in coherent themes. The analysis consisted of three iterations, separately conducted by an urban designer and a computer scientist. In the first iteration, the whole set of POI categories was scanned and initial codes were generated; this was followed

Themes	*Category*	*Greater London*	*Greater Manchester*	*West Midlands*
		Foursquare		
Health harmful food	Fried Chicken	0.31	0.15	0.19
	Fast Food		0.22	0.31
	Wings	0.11		
Faith	Mosque	0.27	0.22	
	Church	-0.18	-0.15	
Non-local cuisines	African	0.32		0.25
	Caribbean	0.37		0.21
	Asian			0.23
	Italian	-0.26	-0.36	-0.25
	Indian	-0.27	-0.17	
	Spanish		-0.20	
	Chinese		-0.22	
Beauty & aesthetics	Dentist's Office	-0.22	-0.21	-0.15
	Nail Salon		-0.17	-0.19
	Salon Barbershop	-0.15		-0.35
Sports	Golf Course	-0.24	-0.28	-0.42
	Cricket	-0.13		-0.23
	Tennis Court		-0.23	
Open spaces	Other Outdoors	-0.15	-0.24	-0.25
	Lake	-0.12	-0.13	
	Campground		-0.22	-0.23
	Field	-0.15	-0.23	
	Playground		-0.22	
	Trail		-0.21	
	Outdoors and Recreation		-0.14	
Bus service	Bus	0.15		0.23
	Bus Station	0.28	0.32	
	Bus Stop	0.18		
		OSM		
Road system elements	traffic signals	0.29	0.25	
	crossing	0.25		
	mini roundabout			0.24

Table 12. Spearman correlation values r_s between Foursquare and OSM categories considered by the thematic analysis and IMD (all results shown are statistically significant, $p < 0.05$).

by merging semantically-related codes into broader themes using relevant urban studies as guidance; finally, identified themes were refined and named. In the end, a total of eight common themes were identified: (derived from Foursquare) *health harmful food*, *faith*, *non-local cuisines*, *beauty & aesthetics*, *sports*, *open spaces*, and *bus service*; (derived from OSM) *road system elements*. Table 12 shows the Foursquare and OSM categories related to each theme, along with the Spearman correlation values for each category within these theme, computed through the Clifford *et al.* method [6] between their *Offering Advantage* and IMD. Note that these correlation values are only valid for the three cities under study; as these three cities belong to the same country, it is not surprising that values for the same POI category are similar across them. However, if we were to apply this method to other cities in the world, the same POI category could bear opposite correlation with deprivation. While correlation findings are expected to differ, the same method could still be applied to other urban contexts, as long as UGC is available. We next briefly elaborate on each of the derived themes; in most cases, to gain confidence in their validity, we mention similar results in the literature.

Health harmful food

We created this theme to include all the Foursquare venues that are related with restaurants selling unhealthy food. These are Fried Chicken, Fast Food and Wings. Those venues are positively associated with neighborhoods with IMD scores above the median. This finding is consistent with some studies in preventive medicine: using qualitative investigations, MacDonald *et al.* found that the higher the density of chain fast-food restaurants, the higher the neighborhood deprivation for England and Scotland [8]. Other studies have been carried out in New Zealand [31] and in the USA [2] and found similar results, thus suggesting the same correlation sign for this theme could be found in cities within these other countries too.

Faith

This theme includes two Foursquare venues: Mosque and Church. However, the two bear opposite correlation with deprivation: mosques tend to have higher concentration in areas with IMD above the median, while churches are more concentrated in areas with IMD below the median. Previous research has shown that there is a link between high concentrations of Muslim residents in London wards and below the median IMD values [5]. This seems to be consistent with part of our finding; however it is also true that Muslims might not live where their places of worship are located. To ascertain this missing point, we studied the relationship between the percentage of Muslims living in a certain area (relative to the total number of religious people in that area) and the *Offering Advantage* for the Foursquare category Mosque. We did so by extracting information from the Census Data 2011 for Greater London at the level of ward.[9] We indeed found a positive correlation between the presence of mosques in a neighborhood and percentages of Muslim residents in it ($r_s = 0.40$, $p < 0.01$). This seems to be congruent with the hypothesis that neighborhoods with Muslim predominance, which are generally associated with above-median IMD values in Greater London [5], have a higher-than-normal number of mosques.

Non-local cuisines

We created this theme to include all Foursquare venues that are related to restaurants but exclude those covering local cuisine (i.e., Pub, Fish & Chips Shop, English Restaurant), as the latter did not bear strong correlation with IMD. Within this broad theme, we identified two sub-themes: one comprising cuisines that, in the cities under consideration, had positive correlation with deprivation (e.g., African, Asian and Caribbean), and one comprising cuisines that, once again for the three cities under consideration, had negative correlation with deprivation (e.g., Italian, Chinese, Spanish, Indian).

Beauty & aesthetics

This theme comprises three Foursquare venues: Dentist, Nail Salon and Salon Barbershop. All these categories are negatively correlated with IMD, suggesting that beauty and aesthetics facilities concentrate in neighborhoods with IMD below the median. We found a reference to this finding for the category Dentist's Office. Previous studies have, in fact, demonstrated that high socioeconomic status is significantly associated with good oral health [23]; our results seem to be congruent with those findings. However, this is an example of a finding that may not generalise to other geographic contexts: for example, previous research has found a link between the prevalence of beauty salons in areas of the USA and their socio-economic deprivation [39]. Note that, although

[9] `http://www.nomisweb.co.uk/census/2011/ks209ew`

the finding itself might not generalise, the methodology we propose to discover whether this is indeed the case remains the very same.

Sports
This theme includes all the Foursquare categories related with physical activity facilities. These are Golf Course, Cricket and Tennis Court. These venues are negatively correlated with deprivation, suggesting that sport facilities tend to be concentrated in areas with low IMD scores. Previous qualitative studies are consistent with this finding: researchers found that golf courses in Australia [14], fitness centers and dance facilities in the USA [32] tend to be more commonly available in wealthier areas.

Open spaces
We create this theme to include all the Foursquare facilities related to public open spaces. These are Lake, Outdoors and Recreation, Playground, Campground, Trail, Field and Other Outdoors. All these categories are negatively correlated with deprivation. Previous studies are congruent with these results and found similar outcomes for the Netherlands [24], in Howard County, USA [13] and for Portland, USA [3].

Bus service
This theme includes the Foursquare categories for Bus, Bus Stop and Bus Station. These venues are positively correlated with deprivation; their concentration in an area is thus signal of higher than average deprivation. This finding is consistent with a recent report issued by Transport for London that shows that, for Greater London, the share of bus trips increases with the decrease of household incomes [44].

Road system elements
This theme includes all the OSM amenity types related with the traffic management system. These are traffic signals, crossings (i.e., elements which guarantee safe pedestrian crossing) and mini roundabouts. All of these OSM amenities are positively correlated with deprivation and therefore tend to be concentrated in areas with IMD scores above the median. Higher-than-normal presence of these elements may be associated with road infrastructures (e.g., junctions, highways, main roads), which might make a neighborhood less attractive to live in.

IMPLICATIONS, LIMITATIONS AND FUTURE WORK

In the previous sections, we have proposed and evaluated a new method that mines urban data obtained from easily accessible UGC datasets (namely, Foursquare and OpenStreetMap) to extract features that correlate with metrics of socioeconomic deprivation at the level of cities' neighborhoods. We have shown that we can use these features to build accurate classifiers of deprivation and that the corresponding results are in line with previous findings.

Implications
This work has both practical and theoretical implications. From a practical standpoint, the suggested method affords the ability to build 'neighborhood profiling' tools that different stakeholders can use for different purposes: for example, residents may use them to decide where to buy a property or rent a flat; visitors may consult them to decide in which hotel or rent-house to stay; and city planners and administrators may use them to analyze and compare what makes a well-off *vs.* a deprived neighborhood in their cities. From a theoretical standpoint, our method can be used by urban designers and social science researchers to advance knowledge in their fields: for example, to understand the relationship between the built environment and deprivation, as it varies across different cities and cultures; and to analyze the relationships between the built environment and deprivation as it varies in response to different processes such as urbanization and gentrification. Note that, when analysing different geographic contexts, findings elicited with our methodology may well differ; however, the methodology itself is generally applicable, and indeed can be applied in different geographic settings to discover these variations.

Limitations
When selecting to apply our proposed method, one needs to take into consideration the following limitations. First, both Foursquare and OSM datasets have geographic and social biases. The two datasets, in fact, do not have a uniform coverage of urban features across space; rather, coverage is concentrated in city centers, thus affording us only a partial picture of what elements are present in areas further away [26]. In terms of social bias, both Foursquare and OSM users belong to a rather specific demographic group (i.e., young, educated, wealthy); one may thus question how representative the data they produce is of what indeed exists in the physical space. When applying our methodology, one should first check for geographic and social biases, for example using the method proposed in [33]; for cities where such effects are large, our methodology is more likely to return invalid results.

The second limitation which ought to be acknowledged concerns the multiple comparison problem. Our method requires the simultaneous test of multiple variables, and this might increase the chance of finding false positive correlations. Statistical control techniques like FDR should be applied, before deciding whether risks of invalid findings are low. In the case studies presented in this work, we applied FDR to estimate such risk; results appear robust when using Foursquare data, especially for the cities of Greater Manchester and West Midlands.

The third limitation has to do with the coarse-grained classification granularity. We have demonstrated clear performance advantage of our method over a plausible baseline and over a state-of-the-art method that classifies urban deprivation in a binary fashion. In the future, to perform finer-grained classifications, researchers might use multi-modal machine learning approaches to combine features derived from multiple datasets. For example, they could combine social media data with image data coming from Google Street Views, to obtain information about the aesthetic capital afforded by many neighborhoods around the world [19, 36].

Finally, for moderate levels of urban deprivation, classification accuracy should be improved, since our classifier performs really well only at the two extremes of the distribution. One way of doing so is to explore metrics other than

Offering Advantage. There might be metrics that capture different notions of a neighborhood's potential offering (e.g., what types of POIs are within walking distance) and that express those notions with alternative mathematical formulations (e.g., graph-based ones [18]).

Future work

Our future work spans two main directions: on one hand, we aim to expand the method so to capture other urban features (e.g., walkability) that pertain the physical layout of neighborhoods. OSM data is particularly well-suited to extract such features, as it comprises not just POI data (as is the case for Foursquare), but also road network information. Examples of features we might computationally capture include average block size and road density, that urban studies have previously linked to indicators of neighborhood well-being [45, 1]. On the other hand, we aim to quantify the applicability of our method to measure deprivation in different geographic contexts, especially in developing countries where accurate and up-to-date indexes of deprivation like IMD are very difficult to build. We note that, in these contexts, UGC data as obtained by social media services like Foursquare is probably of very little value (too sparse); however, OSM penetration has been shown to be high also in developing countries [34], thus suggesting our method could be used in these more challenging contexts too, especially once we have expanded our set of features to comprise those that we can extract from the road network.

Acknowledgments

This work has been carried out thanks to the Engineering and Physical Sciences Research Council (EPSRC) and the Centre for Urban Sustainability and Resilience (USAR) at University College London.

REFERENCES

1. Alberti, M. The effects of urban patterns on ecosystem function. *International Regional Science Review 28*, 2 (2005), 168–192.
2. Block, J. P., Scribner, R. A., and DeSalvo, K. B. Fast food, race/ethnicity, and income: a geographic analysis. *American journal of preventive medicine 27*, 3 (2004), 211–217.
3. Bolitzer, B., and Netusil, N. R. The impact of open spaces on property values in Portland, Oregon. *Journal of environmental management 59*, 3 (2000), 185–193.
4. Braun, V., and Clarke, V. Using thematic analysis in psychology. *Qualitative Research in Psychology 3*, 2 (2006), 77–101.
5. Brimicombe, A. J. Ethnicity, religion, and residential segregation in London: evidence from a computational typology of minority communities. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN 34*, 5 (2007), 884.
6. Clifford, P., Richardson, S., and Hemon, D. Assessing the significance of the correlation between two spatial processes. *Biometrics 45*, 1 (1989), 123–134.
7. Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. The Livehoods Project: utilizing social media to understand the dynamics of a city. In *Proc. of ICWSM*, AAAI (2012).
8. Cummins, S. C. J., McKay, L., Witten, K., and MacIntyre, S. McDonalds restaurants and neighborhood deprivation in Scotland and England. *American Journal of Preventive Medicine 29*, 4 (2005), 308–310.
9. Department of Economic and Social Affairs. World urbanization prospects, the 2011 revision: highlights. Tech. rep., Population Division United Nations, 2011.
10. Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science 328* (2010), 1029–1031.
11. Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., and Davis, E. R. Mapping city lights with nighttime data from the DMSP Operational Linescan System. *Photogrammetric engineering and Remote Sensing 63*, 6 (1997), 727–734.
12. Elvidge, C. D., Imhoff, M. L., Baugh, K. E., Hobson, V. R., Nelsonc, I., Safran, J., Dietz, J. B., and Tuttle, B. T. Night-time lights of the world: 1994-1995. *Photogrammetry and Remote Sensing 56*, 2 (2001), 81–99.
13. Geoghegan, J. The value of open spaces in residential land use. *Land use policy 19*, 1 (2002), 91–98.
14. Giles-Corti, B., and Donovan, R. Socioeconomic status differences in recreational physical activity levels and real and perceived access to a supportive physical environment. *Preventive Medicine*, 35 (2002), 601–611.
15. Grenyer, R., Orme, C. D. L., Jackson, S. F., Thomas, G. H., Davies, R. G., Davies, T. J., Jones, K. E., Olson, V. A., Ridgely, R. S., Rasmussen, P. C., Ding, T., Bennett, P. M., Blackburn, T. M., Gaston, K. J., Gittleman, J. L., and Owens, I. P. F. Global distribution and conservation of rare and threatened vertebrates. *Nature 444* (2006), 93–96.
16. Griffith, D. A., and Paelinck, J. H. *Non-standard spatial statistics and spatial econometrics*. Springer, 2011.
17. Hidalgo, C. A., Klinger, B., Barabási, A. L., and Hausmann, R. The product space conditions the development of nations. *Science 317*, 5837 (2007), 482–487.
18. Hillier, B., and Hanson, J. *The social logic of space*. Cambridge University Press, 1984.
19. Hwang, J., and Sampson, R. J. Divergent Pathways of Gentrification: Racial Inequality and the Social Order of Renewal in Chicago Neighborhoods. *American Sociological Review* (2014).
20. Lacy, S. Foursquare closes $50M at a $600M valuation. Tech. rep., TechCrunch, 2011.
21. Legendre, P. Spatial autocorrelation: trouble or new paradigm? *Ecology 74*, 6 (1993), 1659–1673.

22. Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J. I'm the mayor of my house: examining why people use Foursquare - a social-driven location sharing application. In *Proc. of CHI*, ACM (2011), 2409–2418.

23. Locker, D. Deprivation and oral health: a review. *Community dentistry and oral epidemiology 28*, 3 (2000), 161–169.

24. Luttik, J. The value of trees, water and open space as reflected by house prices in the Netherlands. *Landscape and Urban Planning 48*, 3 (2000), 161–167.

25. Mao, H., Shuai, X., Ahn, Y. Y., and Bollen, J. Mobile communications reveal the regional economy in Cote d'Ivoire. In *Proc. of NetMob* (2013).

26. Mashhadi, A., Quattrone, G., and Capra, L. Putting Ubiquitous Crowd-sourcing into Context. In *Proc. of CSCW*, ACM (2013), 611–622.

27. McLennan, D., Barnes, H., Noble, M., Davies, J., Garratt, E., and Dibben, C. The english indices of deprivation 2010. *London: Department for Communities and Local Government* (2011).

28. Moreno, E. L., Bazoglu, N., Mboup, G., and Warah, R. State of the world's cities 2008/2009 - harmonious cities. Tech. rep., UN-HABITAT, 2008.

29. Noor, A. M., Alegana, V. A., Gething, P. W., Tatem, A. J., and Snow, R. W. Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics 6*, 5 (2008), 81–99.

30. Overseas Development Institute. Briefing paper 44: Opportunity and exploitation in urban labour markets. Tech. rep., 2008.

31. Pearce, J., Blakely, T., Witten, K., and Bartie, P. Neighborhood deprivation and access to fast-food retailing: a national study. *American journal of preventive medicine 32*, 5 (2007), 375–382.

32. Powell, L., Slater, S., Chaloupka, F., and Harper, D. Availability of physical activity-related facilities and neighborhood demographic and socioeconomic characteristics: a national study. *American Journal of Public Health 96*, 9 (2006), 1676–1680.

33. Quattrone, G., Capra, L., and Meo, P. D. Theres No Such Thing as the Perfect Map: Quantifying Bias in Spatial Crowd-sourcing Datasets. In *Proc. of CSCW*, ACM (2015).

34. Quattrone, G., Mashhadi, A., and Capra, L. Mind the map: the impact of culture and economic affluence on crowd-mapping behaviours. In *Proc. of CSCW*, ACM (2014).

35. Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. Tracking Gross Community Happiness from Tweets. In *Proc. of CSCW*, ACM (2012).

36. Quercia, D., Ohare, N., and Cramer, H. Aesthetic Capital: What Makes London Look Beautiful, Quiet, and Happy? In *Proc. of CSCW*, ACM (2014).

37. Quercia, D., and Saez, D. Mining urban deprivation from Foursquare: implicit crowdsourcing of city land use. *Pervasive Computing, IEEE 13*, 2 (2014), 30–36.

38. Quercia, D., Séaghdha, D. O., and Crowcroft, J. Talk of the city: our tweets, our community happiness. In *Proc. of ICWSM*, AAAI (2012).

39. Small, M. L., and McDermott, M. The presence of organizational resources in poor urban neighborhoods: An analysis of average and contextual effects. *Social Forces 84*, 3 (2006), 1697–1724.

40. Smith, C., Mashhadi, A., and Capra, L. Ubiquitous sensing for mapping poverty in developing countries. In *Paper submitted to the Orange D4D Challenge (2013)* (2013).

41. Smith, C., Quercia, D., and Capra, L. Finger on the pulse: identifying deprivation using transit flow analysis. In *Proc. of CSCW*, ACM (2013), 683–692.

42. Storey, J. D., and Tibshirani, R. Statistical significance for genomewide studies. In *Proc. of NAS* (2003), 9440–9445.

43. Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Economic Geography 46* (1970), 234–240.

44. Transport for London (TFL). Travel in London, Supplementary Report: London Travel Demand Survey (LTDS). Tech. rep., 2011.

45. Tratalos, J., Fuller, R. A., Warren, P. H., Davies, R. G., and Gaston, K. J. Urban form, biodiversity potential and ecosystem services. *Landscape and Urban Planning 83*, 4 (2007), 308 – 317.