UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note
RN/12/08

Putting Ubiquitous Crowd-Sourcing into Context

18-09-2012

Afra Mashhadi

Giovanni Quattrone

Licia Capra

Abstract:

Ubiquitous crowd-sourcing has become a popular mechanism to harvest knowledge from the masses. OpenStreetMap (OSM) is a successful example of ubiquitous crowd-sourcing, where citizens volunteer geographic information in order to build and maintain an accurate map of the changing world. Research has shown that OSM information is accurate, by comparing it with centrally maintained spatial information such as Ordnance Survey. However, we find that coverage is low and non uniformly distributed, thus challenging the suitability of ubiquitous crowd-sourcing as a mechanism to map the whole world. In this paper, we investigate what contextual factors correlate with coverage of OSM information in urban settings. We find that, although there is a direct correlation between population density and information coverage, other socio-economic factors also play an important role. We discuss the implications of these findings with respect to the design of urban crowd-sourcing applications.

Putting Ubiquitous Crowd-sourcing into Context

ABSTRACT

Ubiquitous crowd-sourcing has become a popular mechanism to harvest knowledge from the masses. OpenStreetMap (OSM) is a successful example of ubiquitous crowd-sourcing, where citizens volunteer geographic information in order to build and maintain an accurate map of the changing world. Research has shown that OSM information is *accurate*, by comparing it with centrally maintained spatial information such as Ordnance Survey. However, we find that coverage is low and non uniformly distributed, thus challenging the suitability of ubiquitous crowd-sourcing as a mechanism to map the whole world. In this paper, we investigate what contextual factors correlate with *coverage* of OSM information in urban settings. We find that, although there is a direct correlation between population density and information coverage, other socio-economic factors also play an important role. We discuss the implications of these findings with respect to the design of urban crowd-sourcing applications.

Author Keywords

Crowd-sourcing, Volunteered Geographic Information, Socio-Economic Factors

ACM Classification Keywords

H.2.8 Database Management: Database Applications—*Spatial Databases and GIS*

INTRODUCTION

Cities are highly dynamic entities, with urban elements such as businesses, cultural and social Points-of-Interests (POIs), housing, and the like, continuously changing. Maintaining accurate spatial information in these settings has become an incredibly onerous task, rendering some centrally-maintained public datasets obsolete [14]. A solution made possible by the upraise of social media is crowd-sourcing, where user-generated content can be cultivated into meaningful and informative collections, as exemplified by sites like Wikipedia [24]. This form of crowd-sourcing is no longer confined to the Web: equipped with powerful mobile devices, citizens have become surveyors, with council-monitoring applications like FixMyStreet;[1] reporters, with micro-blogging sites such as Twitter;[2] and cartographers, with geo-wikis like Cyclopath[3] and OpenStreetMap.[4]

OpenStreetMap (OSM) is perhaps one of the most successful examples of ubiquitous crowd-sourcing, with currently over 547,270 users, collectively building a free, openly accessible, editable map of the world. OSM exhibits *ubiquitous* features, because of the spatio-temporal nature of the knowledge it gathers (map elements of the changing world). Furthermore, Hecht et al. [9] have shown that the "localness" of participation in repositories of user-generated content with geospatial component is high if the available editing tools make use of GPS, as is the case with OSM (accessible and editable via mobile phone applications). It can thus be assumed that editing urban elements in OSM is done by citizens who have actually visited that location.

The geographic information stored in OSM has been demonstrated to be of high quality, where quality has been mainly measured in terms of positional accuracy. Indeed, OSM's accuracy has shown to sometimes supersede the most reputable geographic datasets, performing especially well in urban areas [6, 13]. However, relying entirely on user-generated content for urban mapping raises concerns, not only in terms of accuracy of the collected information (which, for OSM, is presently high), but crucially in terms of *coverage*. In other words, *what part* of the physical world is being digitally mapped? Studies that looked at who the main contributors of crowd-sourcing systems (i.e., Wikipedia) are have shown these to be a group of predominantly young and educated male [3]; they also reported a large gender gap among editors (87% male vs. 13% female). As the crowd-sourcing user base is not representative of the world population, can we expect the geographic content they contribute to be representative of the whole physical world? To answer this question, we performed a study in the area of Greater London, UK, where OSM was originally created and launched, and where the community of contributors is particularly active. As we shall demonstrate in the paper, OSM map features are *not* uniformly distributed across the city. This raises a fundamental question: what contextual factors contribute to coverage of volunteered geographic information in urban settings?

[1] `http://www.fixmystreet.com/`
[2] `https://twitter.com/`
[3] `http://cyclopath.org/`
[4] `http://www.openstreetmap.org/`

Submitted for review to CSCW 2013.

Answering this question is necessary, so to understand *where* crowd-sourced map information can be relied upon (and crucially where not), with direct implications on the design of applications that rely on having complete and unbiased map knowledge.

In this work, we investigate to what extent various socio-economic factors of urban areas correlate with coverage of crowd-sourced geo-spatial data. Although this research question has been studied extensively in the social sciences [22, 7], it has received limited attention from the ubiquitous computing community. Thus, in this paper, we report on a study that aims at discovering the contextual factors that impact coverage of information in OSM for the city of London, UK. As one can expect, we find that coverage is directly correlated with population density; however, we also find that other socio-economic factors are highly significant.

The rest of this paper is structured as follows: after a brief overview of the state-of-the-art in ubiquitous crowd-sourcing research, we describe the dataset at hand, the metrics we computed, and the methodology we adopted. We then illustrate the results of our analysis, before moving on to the discussion section, where we state the implications of these findings for the design of ubiquitous crowd-sourcing applications. We finally conclude the paper and elaborate on future directions of research.

BACKGROUND AND RELATED WORK

Ubiquitous crowd-sourcing is a form of collective gathering of user-generated content that has seen a massive uptake in recent years, thanks to the combined and wide adoption of mobile technology and social media. A popular example of user-generated content is volunteered geographic information (VGI), such as that maintained by OSM. In order for businesses (e.g., Foursquare) to rely on VGI as opposed to proprietary datasets (e.g., Google Maps), quality of the contributed information must be high. For years, the research community has studied the quality of such information [4], compared to traditional geographical datasets maintained by national mapping agencies, as well as proprietary datasets maintained by commercial companies such as Navteq.[5] The findings show very high accuracy: for example, Haklay et al. [5, 6] measured the positional accuracy of OSM road networks in the UK and found it to be very high (i.e., on average within 6 meters of the position recorded by Ordnance Survey). The authors have also investigated the impact of the number of contributors on positional accuracy, and estimated that high accuracy is achieved when there are at least 15 contributors per square kilometre. Works such as [2, 12] have confirmed these observations for countries like France, Germany and Switzerland. Moving from accuracy to coverage of OSM data, a recent study by Zielstra et al. [25] shows that coverage in Germany sharply decreases as we move away from city centres; Girres et al. [2] also discovered a correlation between the number of OSM objects in an area and number of contributors in that area (i.e., areas with up to three contributors per square kilometre had ten times more contributions than areas with only one contributor, and areas with more than three contributors had up to hundred times more contributions).

A limitation to the studies conducted by the VGI community on OSM is the focus on road networks only. However, the contribution process associated with editing roads and that associated with editing Points-of-Interests, such as restaurants and cafes, differ greatly: indeed, the former is typically done by a selected number of users who have high expertise in both the geography of an area and the editing tools required to digitally represent it; the latter can be performed by any city dweller owning a GPS-enabled smart-phone instead. It is the latter that is most representative of citizen engagement. In this work, we thus focus on eliciting the factors that relate to *coverage* of crowd-sourced POIs in urban areas.

OSM is not the only example of crowd-sourced urban information. For example, Cyclopath is being successfully used to digitally map route information for bicyclists in Minneapolis. The system has been widely studied by the academic community, both in terms of its design and rationale [18], its effectiveness [20] and in terms of user's participation and behavioural analysis [19, 15, 16]. In [19], for example, the authors investigated the techniques and motivations that lead to an increased amount of volunteered geographic information in Cyclopath. They found that visually highlighting contribution opportunities and asking users to work on an area that they are mostly familiar with, lead to better coverage. Similarly, [18, 20] discovered that cyclists were interested in sharing their expertise with each other, to cover gaps in terms of missing routes in the geo-wiki, thus increasing coverage of the crowd-sourced information.

These studies offered valuable insights into the *motives* behind user's participation and the impact they have on urban crowd-sourcing. Another important aspect is understanding the *contextual* factors that may affect crowd-sourcing coverage. This line of research has been explored extensively in Wikipedia, where contextual factors of contributors have been analysed in relation to coverage. For example, [10] studied gender imbalance in Wikipedia, and reported on how topics of particular interest to females were substantially less covered than topics of specific interest to males. In [8], the indegree summation (i.e., number of inlinks per article in the Wikipedia Article Graph) on 15 different language editions of Wikipedia was analysed; their findings suggest that population is not the most important factor to be considered, and other factors such as *fluency* in languages are more strongly correlated with indegree instead. They conclude that, when developing technologies to rely upon community maintained repositories, contextual factors of the contributors, such as language and culture, must be carefully examined. This has been done in other user-generate content datasets too; for example, [17] proposes a machine learning technique for estimating location and gender of Flicker users based on the tags they associate to the content they produce.

The work we present in this paper falls into the stream of research that aims to understand the relation between contex-

[5] http://www.navteq.com/

tual factors and coverage of user-generated content. More precisely, we focus our attention on OSM, an example of ubiquitous crowd-sourcing, where content has a distinct spatio-temporal nature. In this domain, we aim to understand the impact that *urban factors*, such as population density, distance to the city center, poverty, and the like, have on OSM coverage. We delve into this study next.

RESEARCH METHODOLOGY

Dataset Description

We begin our study with a detailed description of the crowd-sourcing dataset at hand, that is, OpenStreetMap. The dataset is freely available to download and contains the history of all edits (since 2006) on all spatial objects performed by all users. In OSM jargon, spatial objects can be one of three types: *nodes*, *ways*, and *relations*. Nodes are single geospatial points, defined using latitude/longitude coordinates, and they typically represent POIs; ways consist of ordered sequences of nodes, and mostly represent roads (as well as streams, railway lines, and the like); finally, relations are used for grouping other objects together, based on logical (and usually local) relationships (e.g., administrative boundaries, bus routes).

For the purpose of this study, we restricted our attention to nodes only. In particular, as our choice of sampling strategy, we focused on those that represent urban elements commonly interpreted as *leisure* POIs, such as cafes, restaurants, pubs and bars. These are the categories that are most common to mobile applications such as MyCityWay,[6] Google HotPot[7] and Foursquare,[8] which are used by city dwellers to navigate the urban landscape. To ensure we are considering genuine crowd-sourcing contributions, and not those made by bots (i.e., mass imports), we have eliminated from the dataset those users who performed an excessive number of edits in a very short time (i.e., those who edited more than 40 POIs in a single *changeset* session in OSM, with the threshold of 40 chosen after manual inspection of the per-user edit distribution). Finally, we focused on the area of Greater London, UK, which is an example of urban city with many administrative districts with different socio-economic factors (as we shall present later). The resulting crowd-sourcing dataset consists of 818 users, editing 9,573 POIs by means of 19,139 edits overall.

In order to compute coverage of OSM, we required *(i)* a benchmark against which to compare OSM POIs and *(ii)* a matching algorithm to map OSM POIs to those in the benchmark dataset.

Benchmark. We required a ground-truth dataset, containing all POIs physically present in each chosen area. For this purpose, we selected Navteq, the leading global provider of maps and location data, covering not only roads but also millions of POIs of varying nature, from restaurants to hospitals and gas stations. Being a commercial service, Navteq's primary objective is to ensure the highest level of accuracy of its data (the information contained there is factually correct and up-to-date).

[6] `http://www.mycityway.com`
[7] `http://www.google.com/hotpot`
[8] `https://foursquare.com`

Amenity	*Perc. in OSM*
Post box	18%
Nightlife	15%
Place of Worship	6%
Restaurant	6%
Bicycle Parking	6%
School	5%
Cafe	4%
Other	40%

Amenity	*Perc. in Navteq*
Restaurant	12%
Vehicle Repair	8%
School	8%
Clothing Store	6%
Nightlife	6%
Cafe	5%
Grocery Store	3%
Other	52%

Table 1. Amenity Distribution in OSM and Navteq

Table 1 reports on the most popular amenity categories in OSM and Navteq separately. It is worth noting that, while OSM also deals with objects that are of interest to the community, such as post boxes (18%) and bicycle parking (6%), Navteq is primarily concerned with commercial entities instead, such as restaurants (12%) and clothing stores (6%). In this work, we restrict our attention to what we call *leisure* POIs, to indicate those categories which have a presence in both OSM and Navteq.

POI Matching Algorithm. To be able to measure coverage, we first need to relate POIs in OSM with the same POIs in the ground-truth dataset in an automatic way. In both OSM and Navteq, a POI is defined as a tuple: $poi = \langle name, (lat, lon)\rangle$, where $name$ is the POI's name, and (lat, lon) are the coordinates determining its geographical position. We then define and quantify, for each POI in OSM, two measures: *geographic error* and *lexicographic error*. More precisely, let poi_x be a single POI, and POI_x the set of all POIs, with x being either the OSM dataset or the ground-truth dataset (to which we will refer, for convenience, simply as gt). We thus state that poi_{osm} is *equivalent* to poi_{gt} if *both* their geographic distance *and* lexicographic differences are below some specific thresholds. The geographic distance $geo_{Err}(\cdot,\cdot)$ is computed as the Euclidean distance between the two points, while the lexicographic difference $lexical_{Err}(\cdot,\cdot)$ is computed as the Levenshtein distance between the POI names normalized between [0,1] by the length of the POI names.

To determine suitable thresholds to use in our POI matching algorithm, we proceeded as follow: we first considered a subset of 100 POIs from OSM, computed geographic and lexicographic distance to all ground-truth POIs, and determined the 'closest match' for each of these. We then manually inspected which of these were indeed matches, and which were not. Based on this inspection, we empirically derived a threshold of 100 meters for the geographic distance, and a threshold of 0.33 for the lexicographic difference. To further validate these choices, we ran the matching algorithm using these thresholds; upon completion, we selected a small (different) subset of 30 POIs, and manually inspected the correctness of the matches. We found that 97% of these had been correctly matched, while only 3% were not. For illustrative purposes, Table 2 presents some examples of POIs that our matching algorithm correctly relates (first three), despite some lexicographic and geographic error; it also illus-

OSM name	*ground-truth name*	geo_{Err}	$lexical_{Err}$	*Equiv.*
Rondhouse	Rondhouse	0 m	0	Yes
The Green Gate	Green Gate	33 m	0.28	Yes
Smollenskys Bar	Smollensky's	48 m	0.33	Yes
Eardley Arms	Eardley Garage	145 m	0.29	No
Whittington Stone	Whittington NHS	180 m	0.29	No

Table 2. Some example of POIs in OSM and ground-truth dataset with different values of geographic distance and lexicographic difference

trates two examples of correct mismatches (the POIs in OSM are pubs, while those in Navteq are a garage and a hospital respectively).

Metric. Based on the above mapping, we have evaluated coverage of OSM POIs for Greater London as:

$$coverage = \frac{\#(\{\text{POIs in OSM}\} \cap \{\text{POIs in Navteq}\})}{\#\{\text{POIs in Navteq}\}}$$

with $coverage \in [0, 1]$. The higher the coverage, the higher the extent to which the ground-truth POIs are also present in OSM.

Contextual Factors of OSM

Our hypothesis is that there is a strong relationship between socio-economic characteristics of an urban area and the level of coverage that can be expected of this area by means of volunteered contributions. To validate this hypothesis, we focus on OSM coverage for London at a finer level of granularity than the city level, that is, the level of wards. We have chosen this level of granularity as wards are the smallest regions defined by local authorities in London.[9]

London presently comprises 600 wards. Figure 1 illustrates the frequency distribution of ground-truth POIs at ward level. As it can be seen from the head of the distribution, there are many wards with less than 5 POIs, and a long tail of a few wards with many POIs. To avoid biased analysis due to sparsity of this data, we have considered only wards that have 5 or more POIs. In so doing, we disregarded 120 wards, analysing 480 of the original 600. For each ward, we have collected the following contextual factors:

Population. Using UK Census 2011 data published by the National Statistics Office,[10] we have information about population at ward level. Previous studies of OSM coverage for road networks have revealed a correlation between the number of contributors in an area and the number of OSM objects digitally mapped in that area [2]. We have thus selected population as an attribute for investigation in this study, as it can give us an expectation of contributions per area. Although higher population density does not directly translate into higher number of contributors, we may expect more contributors per unit area to exist in denser areas. The hypotheses we thus want to test are: *(i)* the higher the *population density* of an area (that is, population divided by ward size), the higher the coverage; and *(ii)* the higher the population per POIs in an area (that is, population divided by number of POIs), the higher the coverage.

[9] http://data.london.gov.uk/datastore/package/ward-profiles-2011

[10] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/index.html

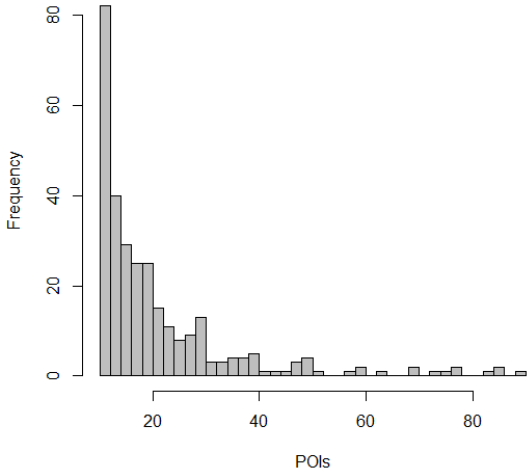

Figure 1. Frequency Distribution of POIs at Ward Level

Poverty. Analysing the relationship between poverty of an area and coverage is important, as it may reveal the impact that (lack of) technology adoption (e.g., use of smartphones and Internet), as well as (lack of) available leisure time, has on it. In this regard, UK Census data contains information about the Indices of Multiple Deprivation (IMD). IMD are a set of indicators, published by the UK Office for National Statistics, measuring deprivation of small geographic areas known as Lower-layer Super Output Areas (LSOA) in England. IMD consist of seven domain indicators. The one we are interested in this study is the Income Deprivation Index, that measures the number of people claiming income support, child tax credits, or asylum; we refer to this factor as *poverty* henceforth. The hypothesis under test is that poverty of an area is negatively correlated with digital mapping of its POIs. This is another important aspect to look into, if we are to rely on volunteered mapping information alone, as it may reveal where gaps arise, thus enabling intervention via contingency plans.

Dynamic Population. While the previous two factors capture 'static' characteristics of the residents of an area, they do not reveal much about the actual pulse of the city, that is, where city dwellers (i.e., the potential contributors of ubiquitous crowd-sourced information, be them residents or tourists) spend time. We thus add a dynamic attribute based on Foursquare check-ins, which we refer to as *dynamic population*.

We acknowledge that Foursquare and OSM share commonalities: neither represents fully the urban population, with a bias towards young, educated and wealthy people; furthermore, the type of content they gather has a common spatio-temporal nature. However, despite these commonalities, we do not expect the behaviour of the crowds that contribute to these systems to be the same: this is because, in Foursquare, users contribute data (check-ins) to

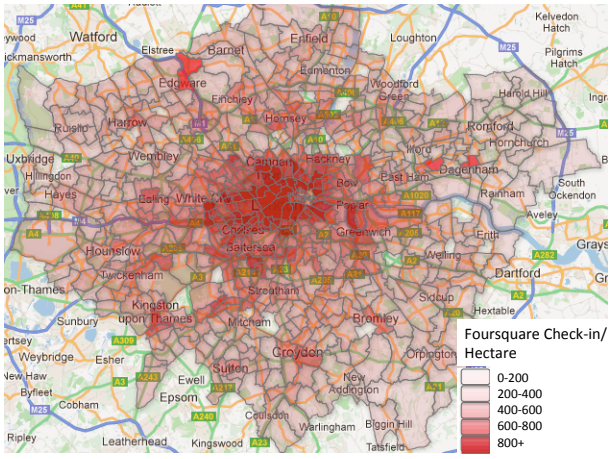

Figure 2. Choropleth map of Foursquare check-ins - Darker wards have higher check-in density

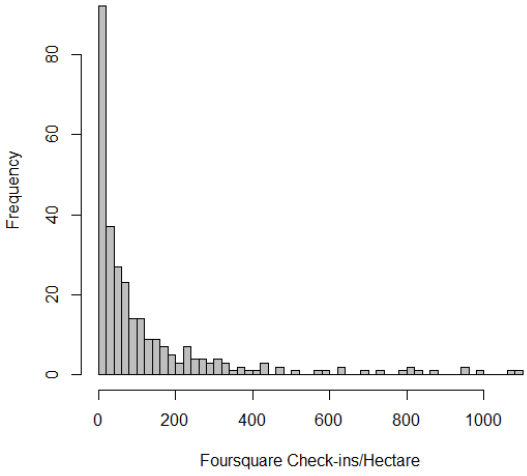

Figure 3. Frequency Distribution of Foursquare Check-in Density

show/share their location and social life with their friends, reflecting an *egocentric* behaviour, whereas OSM editors contribute in order to improve the existing map, thus exhibiting a *community* behaviour. We are thus interested in studying what dynamic population can reveal with respect to coverage. We measure dynamic population based on the last four years of Foursquare activity in London, computing the density of check-ins per ward (i.e., the total number of check-ins in a ward divided by its size); Figure 2 shows a choropleth map of such density distribution across all wards in London: the darker the ward, the higher the density of check-ins. Similarly Figure 3 illustrates the histogram approximating the frequency distribution of density of Foursquare check-ins at ward level: as shown, there are many wards with very low check-ins density, and a long tail of wards with higher check-ins density.

Distance to the Closest Poly-centre. The last factor we consider is the distance from where the social and economic activities happen. Previous studies on OSM have shown

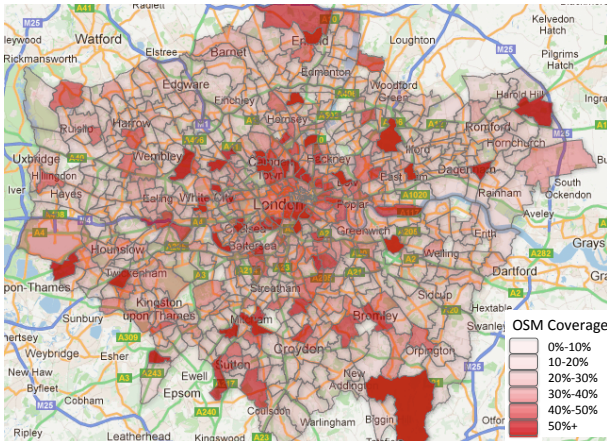

Figure 4. Choropleth map of OSM coverage for Greater London- Darker wards have higher coverage

that road coverage decreases when moving away from the city centre [25]. Similarly, we are interested to examine the effect of distance from the city centre on coverage. However, in metropolitan cities there is not just one centre but multiple urban hubs [1]. Specifically, a recent study [21] has found that London has 10 different polis. In this work, we thus computed the Euclidean distance from the geographic centre point of each ward to the geographic centre point of each of the 10 polis. We then used the shortest distance as our 'distance from the centre' factor, and tested the hypothesis that the closer to the centre, the higher the coverage.

RESEARCH RESULTS

This section reports on the results of our analysis. We first considered the area of Greater London as a whole, for which we found coverage to be 0.35. However, this single aggregate value does not reveal much in terms of what areas of London are being digitally mapped. Figure 4 illustrates the choropleth map of London's coverage, where each tile represents a ward. As shown, coverage is non-uniformly distributed across the city. Previous studies on coverage of OSM for road networks have revealed that distance from the city centre is inversely related to coverage [25]; although at a first approximation a similar pattern seems to emerge for POIs too (i.e., the further away we move from the city centre, the worse the coverage), we can also identify various suburban areas with high coverage instead. We thus hypothesise that distance from the city centre cannot fully explain coverage.

Figure 5 further shows the histogram approximating coverage distribution at ward level. As shown, there are many wards where coverage is very low (≈ 0), and a few wards where coverage is quite high (≈ 0.6) instead. We now proceed to analyse what contextual factors contribute to this distribution. The factors we are interested in are those listed in the previous section, that is: population density, population per POI, poverty, dynamic population, and distance from the closest poly-centre. To quantify the extent to which coverage is related to such parameters, we proceeded in two steps:

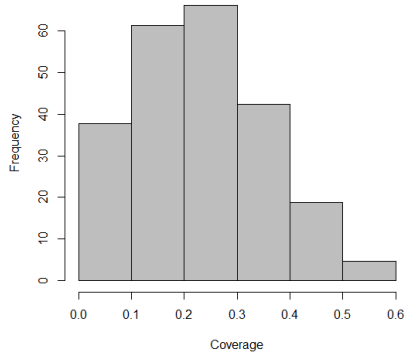

Figure 5. Frequency Distribution of Coverage

Factor	β	R^2	*p-value*
Population Density	0.075	0.10	***
Population per POI	0.005	0.00	
Poverty	-0.021	0.01	*
Dynamic Population	0.090	0.15	***
Distance from the Nearest Poly-centre	-0.085	0.13	***

Table 3. β Coefficient, Multiple R^2 and p-value of Single Linear Regression Models of Coverage on Socio-Economic Factors at Wards Level (p-value significance. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 ' ' 1)**

first, we used single linear regression, considering one parameter at a time as independent variable, and analysed how coverage varies with it. Second, we applied a multiple regression model, so to control for the various parameters at play simultaneously. In all models, all our parameters followed normal distribution, with the exception of dynamic population for which we computed the natural logarithm. Finally, in order to easily interpret the β coefficients of the regression models, we computed the z-scores of each parameter; in so doing, β coefficients indicate the increment in coverage for one unit of standard deviation increment of the corresponding parameter.

We begin analysing the results of the five single linear regression models independently. For each such model, Table 3 reports *(i)* the β *coefficient*, representing the independent contribution of each factor to coverage, *(ii)* R^2, indicating how well each regression model fits the data, and *(iii)* p-value, indicating the significance level of each presented result.

Population Density

We first focus our attention on population density, i.e., the number of people over the number of hectares of each ward. Our intuition is that citizens care most about the area where they live, thus being actively involved in digitally mapping their space. As a consequence, we expect that wards with higher coverage are those where population density is particularly high. Table 3 indeed confirms that population density is positively correlated with coverage ($\beta = 0.075$, $R^2 = 0.10$, p-value < 0.001). In particular, the β coefficient of 0.075 indicates that an increment in population density of 50 people per hectare (i.e., of one unit of standard deviation) would improve coverage of that ward by 0.075. If we consider the distribution of coverage of Greater London (as shown in Figure 5), this increment corresponds to 25% increase in coverage for the average case. The R^2 value is however fairly low, suggesting that a regression model purely based on population density does not fully explain the residual between actual and expected coverage.

To further understand the relation between coverage and population density, the box plot of Figure 6 shows how coverage varies with the change in population density. In particular, the plot for each range of population density presents: a bin graphically depicting the smallest observation, the lower quartile, the median, the upper quartile, and the largest observation for that range of population density. The circles in the figure are the observations that are considered to be outliers. From Figure 6, we can thus see the effect that population density has on coverage (as one grows, so does the other). It is worth noting that there exist very few outliers, thus confirming the validity of our results (i.e., positive correlation between population density and coverage).

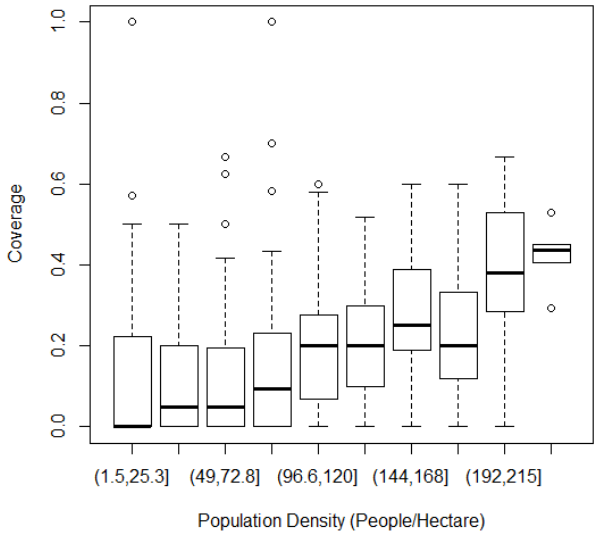

Figure 6. Coverage vs. Population Density

Population per POI

We now move our attention to population per POI, that is, the number of people over the number of (ground-truth) POIs for each ward. The hypothesis we examine here is whether having more people per POI in an area means better coverage of the information in that area. If so, we could then aim to identify a minimum number of people per POI that is required to expect the POI to be mapped, as done in [6] for roads. Interestingly, our analysis (Table 3) reveals otherwise ($\beta = 0.005$, $R^2 = 0.00$, p-value > 0.05). The box plot of Figure 7 depicts variation of coverage with regards to population per POI; as shown population per POI bears no correlation with coverage. In other words, having a higher number of residents per POI does not translate into those POIs being mapped.

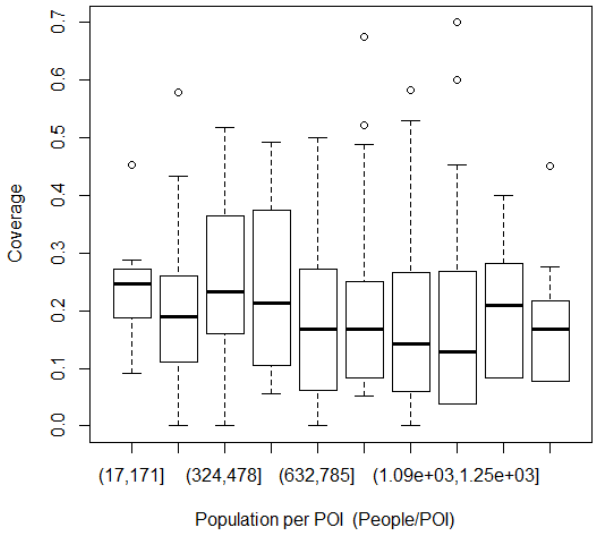

Figure 7. Coverage vs. Population per POI

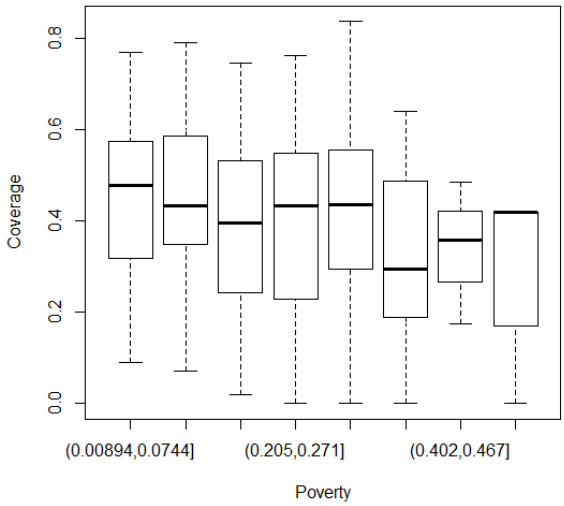

Figure 8. Coverage vs. Poverty

Poverty

We next examine how poverty of an area plays a part in that area being mapped. Studies such as [10] have revealed the contributors of crowd-sourcing to be predominantly a group of young, educated and wealthy males. We thus hypothesise that poverty of an area (measured as deprivation of its residents) is negatively correlated to coverage of that area. Table 3 confirms that poverty of an area has a (weak) negative correlation (β=-0.021) with coverage; that is, a decrement of one unit standard deviation of poverty in a ward would improve coverage of that ward by 0.021 (this increment corresponds to 7% increase in coverage for the average case). Note that $R^2 = 0.01$ is significantly lower than that found for other factors such as population, suggesting that, although significant, poverty itself is only a secondary factor in explaining coverage residual, as computed via linear regression. This is confirmed also by Figure 8 which displays how coverage changes with variations in poverty level. We will return to this point when considering all factors together in a multiple regression model.

Dynamic Population

We now turn our attention to dynamic population. Our hypothesis is that the higher the number of check-ins/visits in an area is, the better mapped such area will be. One may wonder whether poverty of an area and dynamic population of an area (measured as density of Foursquare check-ins) are surrogate of each other, the idea being that poorer areas attract fewer people, while richer areas are expected to attract more businesses and thus more visitors too. However, by performing a correlation analysis between poverty and dynamic population, we discovered the two to be non correlated; in other words, there are areas in London whose residents are income-deprived (e.g., Camden and Hackney) and yet attract large crowds.

Table 3 confirms that dynamic population is highly and significantly correlated with coverage ($\beta = 0.090$, $R^2 = 0.15$, and p-value < 0.001), with β and R^2 values higher than those computed for previous factors. The box plot of Figure 9 also confirms that dynamic population of an area has a positive correlation with coverage.

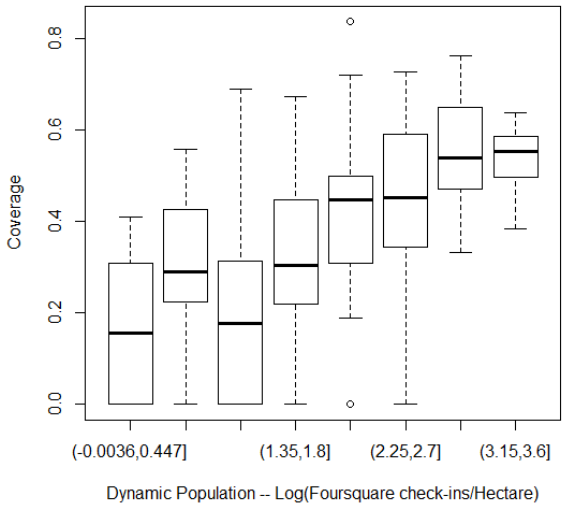

Figure 9. Coverage vs. Dynamic Population

Distance from the Closest Poly-centre

We now turn our attention to the last factor under examination, that is distance to the closest poly-centre. Our hypothesis is that the closer a ward is to the nearest poly-centre of London, the better its coverage. This intuition is confirmed by Table 3, which shows that distance from the closest poly-centre is inversely correlated with coverage ($\beta = -0.085$,

$R^2 = 0.13$, p-value < 0.001). In particular, the β coefficient of -0.085 indicates that a decrement of 5km in distance from the closest poly-centre (i.e., of one unit of standard deviation) would improve coverage of that ward by 0.085 (this increment corresponds to 28% increase in coverage for the average case). Similarly to what we noted before for population density, the R^2 value of a single linear regression model based on distance is relatively low, suggesting that distance from the closest poly-centre does not fully explain the residual between actual and expected coverage.

The box plot of Figure 10 shows how coverage varies as one moves further away from the closest poly-centre: note that, for short distances from the nearest poly-centre, correlation with coverage is indeed rather high; however, as distance increases, this correlation weakens considerably. We will return to this observation in the next section.

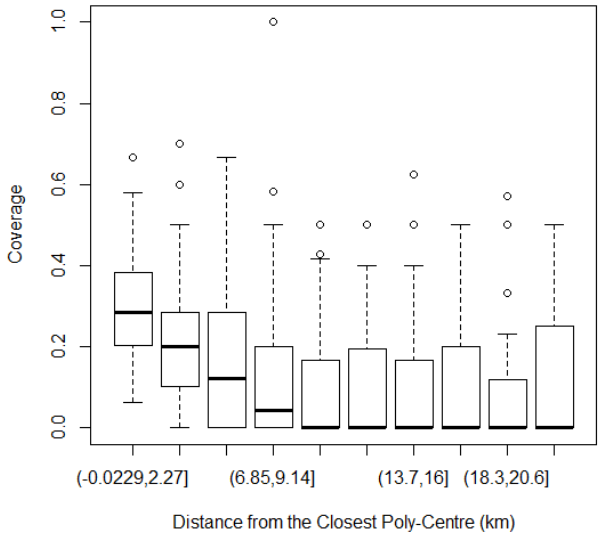

Figure 10. Coverage vs. Distance to the Nearest Poly-centre

Understanding Mediating Influence

Although the previous single predictor models afford us valuable insights into the relations at play between each such variable independently and coverage, in practice we expect coverage to depend on these factors as a whole. We thus need to analyse these parameters together, and understand the relative importance of each of them while controlling for others. We do so by means of a multiple linear regression model.

Table 4 presents the results of such model, reporting β coefficients for each factor and their level of significance, along with multiple R^2. As shown, dynamic population, distance from the nearest poly-centre and population density are the dominant factors, with higher contribution weight and lower p-values. This analysis confirms that population per POI does not contribute to coverage of an area. Finally, the low β coefficient and the high p-value associated with poverty confirms that the correlation between poverty and coverage (as presented in Table 3) is of secondary importance when we consider the mediating influence of all other factors.

Factor	β	*p-value*
Population Density	0.027	*
Population per POI	0.010	
Poverty	-0.007	
Dynamic Population	0.054	***
Distance from the Nearest Poly-centre	-0.031	*
Multiple R^2	0.17	***

Table 4. β Coefficient, Multiple R^2 and p-value of Multiple Linear Regression Models of Coverage on Socio-Economic Factors at Wards Level (p-value significance codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 ' ' 1)**

If we now compare the multiple R^2 value for the multiple regression model (Table 4, $R^2 = 0.17$) with the R^2 value for the best fitting single regression model (Table 3, $R^2 = 0.15$ for dynamic population), we observe only a marginal improvement in terms of model fit. There may be two reasons for this: on one hand, the factors we examined in this work only partly capture the facets of urban context that relate to coverage; future work is required to examine other aspects not included so far. On the other hand, there might be interactions between the factors under study that a simple multiple linear regression model does not capture. In order to evaluate the extent of the impact of these interactions we also considered a multiple regression model with bilinear interactions across all pairs of predictors. We found that the multiple R^2 value of the model with interactions is 0.24 with p-value < 0.001; this means that the model with interactions fits 41% better than the model without, revealing that the effect of interactions between our socio-economic parameters is not negligible. We do not delve further into modelling interactions in this paper; however, we note that, as one moves from understanding the relevance of contextual factors on coverage (i.e., the goal of this work) into building predictor models of coverage growth, such interactions should be explored further (for examples, by means of non-linear models such as SVM).

So far we have attempted to build a model that explains coverage in terms of socio-economic factors, while looking at the area of Greater London as a whole. However, London is a large and complex metropolitan city, and one may wonder whether different regression models should be built and analysed for different sub-areas instead, with the expectation that the same predictors would play a rather different role in such sub-areas. We did so by dividing Greater London in two: Inner London and Outer London, as depicted in Figure 11.[11] The distinction comes from the London Government Act 1963[12] where Inner London is defined as the richest area in Europe, albeit widespread poverty towards the East and South.

We built two multiple regression models for Inner and Outer London separately; we used the model without interactions so to afford direct interpretation of the β parameters with those derived for Greater London as a whole (Table 4). Results for Inner London are reported in Table 5; results for Outer London are aligned with those for the whole of Greater

[11] This image has been taken from `http://wikitravel.org`
[12] `http://www.legislation.gov.uk/ukpga/1963/33`

Figure 11. Inner and Outer London

Factor	β	*p-value*
Population Density	0.013	
Population per POI	0.024	
Poverty	-0.028	*
Dynamic Population	0.008	
Distance from the Nearest Poly-centre	-0.380	***
Multiple R^2	0.19	***

Table 5. β Coefficient, Multiple R^2 and p-value of Multiple Linear Regression Models of Coverage on Socio-Economic Factors for Inner London (p-value significance codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 ' ' 1)**

London and thus not repeated in the interest of space. Note that, while dynamic population and population density were primary factors in relation to coverage when looking at the whole of Greater London, they become secondary factors when we focus on Inner London instead (their β values are lower, and their p-values higher compared to those in Table 4). For Inner London, it is poverty that now shows correlation with coverage (higher β value and lower p-value compared to those reported in Table 4). As an example, we considered two wards in Inner London, one in Chelsea (just north of the river) and one in Battersea (just south of the river, opposite Chelsea). We found that the former has low poverty and high coverage, while the latter has high income deprivation and much lower coverage. Note also that, when focusing on Inner London, distance from the nearest poly-centre is much more strongly related to coverage ($\beta = -0.380$) than when looking at Greater London as a whole ($\beta = -0.031$) . This insight is in accordance with Figure 10, which highlights how distance matters on short and medium length, but less so as one moves away from the center.

DISCUSSION

Limitations

In the previous section we have shown results of an investigation into the relationship between socio-economic factors of an area and the coverage of its POIs in OpenStreetMap. A number of limitations have to be highlighted in relation to the findings previously reported. First, our findings are valid for London, but cannot be directly translated to other cities. We chose to study London because it is an example of a large and complex metropolitan setting, and also because of the rich set of information about this city that is freely available for investigation: being the birth city of OSM, it has a large community of active contributors; furthermore, details of the socio-economic status of its administrative regions is available at a very fine level of granularity. While we cannot expect the findings reported in the previous section to hold true for other cities, the general approach we have presented can be followed to understand what contextual features correlated with coverage in ubiquitous crowd-sourcing domains. For example, one may analyse the extent to which factors such education are correlated with coverage in cities in the developing countries, where there exists a much bigger gap between different groups of the society than in London.

A second limitation relates to the choice of POIs that we have examined (leisure POIs, such as cafes, restaurants, pubs and bars). Our findings cannot be generalised to the mapping of other spatio-temporal information, as it may take place during disaster recovery efforts [26].

Finally, we used census data released in 2011 by the UK Government as measure of population and wealth. This data is valuable, but limited in that it only offers aggregate values per ward. Should we have been in possession of further information, such as wealth distribution and standard deviation within a ward, we could have delved into a more fine-grained assessment of the relationship between these variables and coverage.

Implications

How does the study reported in this paper affect the development of urban crowd-sourcing applications? As our results have highlighted, coverage of VGI in OpenStreetMap varies depending on a variety of contextual factors, in particular distance from the center, dynamic population and population density. Furthermore, in large metropolitan cities like London, the relative importance of each such factor may vary when looking at different geographic clusters; for example, in Inner London coverage is strongly related to poverty but not to dynamic population. Understanding the contextual factors that relate to coverage is important for developers of ubiquitous crowd-sourcing applications, so they can better engineer one. For example, a variety of incentive schemes, spanning from financial rewards to gamification (e.g., in the form of competitions or mapping parties) to location-based social network features [23, 11] can be planned, so to nudge the crowds toward mapping areas that would otherwise be naturally neglected (e.g., because far from the city center, or because they are poor areas within the center).

Understanding the contextual factors of the areas being mapped is only one aspect that developers need to consider in building successful ubiquitous crowd-sourcing applications. Two further aspects require investigation: on one hand, understanding the characteristics of the crowd that the application attracts (for example, locals vs. visitors), and on the other hand the characteristics of the urban objects that such

crowds actually map (for example, services as opposed to leisure POIs). Both directions deserve future investigation.

CONCLUSION AND FUTURE WORK

The study presented in this paper has shown that coverage in OSM, a ubiquitous crowd-sourcing dataset, is non-uniformly distributed across the city. Different contextual factors, including population density, dynamic population, distance from the center and poverty are correlated with information coverage. Raising awareness of the factors that correlate with (lack of) coverage is a first step towards planning interventions, such as developing incentives to nudge the community to take part in a more guided crowd-sourcing act (e.g., to geo-map areas that would otherwise be neglected). Being aware of the contextual factors that affect coverage of crowd-sourced urban information is important for end-users too, so to understand where they can rely on the crowd-sourced information (the risk, in fact, is to make decisions based on partial and biased information).

We are continuing the work started in this paper along two main directions. On one hand, having analysed what contextual factors correlate with coverage in OSM, the next step is to study the crowd-sourcing process as it happens over time. The aim is to build dynamic models that leverage the previously elicited parameters to accurately *predict* what areas will be covered and, crucially, what areas will not, so to direct resources towards targeted interventions.

On the other hand, we are looking at the crowd-sourcing process from a contributors perspectives, rather than from a spatial one. A study focused on OSM contributors will enable us to understand what human factors (both static, such as age and gender, and dynamic, such as mapping patterns) contribute to coverage, and of what type of information. In so doing, we aim to offer a better understanding of the sustainability of crowd-sourcing as a means to gather information about our changing world.

REFERENCES

1. S. Brunn, J. Williams, and D. Zeigler. *Cities Of The World: World Regional Urban Development*. Rowman & Littlefield Publishers, 2003.
2. J. Girres and G. Touya. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4):435–459, 2010.
3. R. Glott, P. Schmidt, and R. Ghosh. Wikipedia Survey–overview of results. *United Nations University: Collaborative Creativity Group*, 2010.
4. M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
5. M. Haklay. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703, 2010.
6. M. Haklay, S. Basiouka, V. Antoniou, and A. Ather. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus Law to Volunteered Geographic Information. *Cartographic Journal, The*, 47(4):315–322, 2010.
7. E. Hargittai and E. Litt. The tweet smell of celebrity success: Explaining variation in twitter adoption among a diverse group of young adults. *New Media & Society*, 13(5):824–842, 2011.
8. B. Hecht and D. Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the 4th International Conference on Communities and Technologies*, pages 11–20. ACM, 2009.
9. B. Hecht and D. Gergle. On the localness of user-generated content. In *Proceedings of the 13th International Conference on Computer supported cooperative work*, pages 229–232. ACM, 2010.
10. S. T. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant, L. Terveen, and J. Riedl. WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 1–10. ACM, 2011.
11. D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. 2011.
12. I. Ludwig, A. Voss, and M. Krause-Traudes. A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing Geoinformation Science for a Changing World*, 1(2):65–84, 2011.
13. A. Mashhadi, G. Quattrone, L. Capra, and P. Mooney. On the Sustainability of Urban Crowd-sourcing for Maintaining Large-scale Geospatial Databases. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration*. ACM, 2012.
14. I. Masser. *Governments and Geographic Information*. Taylor and Francis, London, 1998.
15. K. Panciera, R. Priedhorsky, T. Erickson, and L. Terveen. Lurking? Cyclopaths? A Quantitative Lifecycle Analysis of User Behavior in a Geowiki. In *Proceedings of the 28th International Conference on Human factors in computing systems*, pages 1917–1926. ACM, 2010.
16. K. A. Panciera, M. Masli, and L. G. Terveen. "How should i go from _ to _ without getting killed?": Motivation and Benefits in Open Collaboration. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 183–192, 2011.
17. A. Popescu and G. Grefenstette. Mining user home location and gender from Flickr tags. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. 2010.

18. R. Priedhorsky, B. Jordan, and L. Terveen. How a Personalized Geowiki Can Help Bicyclists Share Information More Effectively. In *Proceedings of the 3rd International Symposium on Wikis and Open Collaboration*, pages 93–98, 2007.

19. R. Priedhorsky, M. Masli, and L. Terveen. Eliciting and Focusing Geographic Volunteer Work. In *Proceedings of the 13th International Conference on Computer Supported Cooperative Work*, pages 61–70. ACM, 2010.

20. R. Priedhorsky and L. Terveen. The Computational Geowiki: What, Why, and How. In *Proceedings of the 11th International Conference on Computer Supported Cooperative Work*, pages 267–276. ACM, 2008.

21. C. Roth, S. M. Kang, M. Batty, and M. Barthlemy. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical fFows. *PLoS ONE*, 6(1), 01 2011.

22. J. Schradie. The Digital Production Gap: The Digital Divide and Web 2.0 Collide. *Poetics*, 39(2):145–168, 2011.

23. Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, and A. Kaltenbrunner. The Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. 2012.

24. J. Voss. Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, pages 24–28, 2005.

25. D. Zielstra and A. Zipf. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In *Proceedings of the 13th International Conference on Geographic Information Science*, 2010.

26. M. Zook, M. Graham, T. Shelton, and S. Gorman. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical and Health Policy*, 2(2), 2010.