The Impact of Society on Volunteered Geographic Information: The case of OpenStreetMap

Afra Mashhadi, Giovanni Quattrone and Licia Capra

Abstract Volunteered Geographical Information (VGI) has been extensively studied in terms of its quality and completeness in the past. However, little attention is given to understanding what factors, beyond individuals' expertise, contribute to the success of VGI. In this chapter we ask whether society and its characteristics such as socio-economic factors have an impact on *what* part of the physical world is being digitally mapped. 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. To answer the above questions, we study over 6 years of crowd-sourced contributions to OpenStreetMap (OSM) a successful example of the VGI paradigm. We measure the positional and thematic accuracy as well as completeness of this information and quantify the role of society on the state of this digital production. Finally we quantify the effect of social engagement as a method of intervention for improving users' participation.

Key words: VGI, Completeness, OpenStreetMap, Socio-Economic Factors

1 Introduction

The advent of Big Data alongside the availability of smartphones with capabilities such as positioning services (GPS) has enabled a new era of Volunteered Geographic Information (VGI). From collaborative mapping to 3D modelling of spatial

Afra Mashhadi
Bell Laboratories, e-mail: afra.mashhadi@alcatel-lucent.com

Giovanni Quattrone
University College London e-mail: g.quattrone@cs.ucl.ac.uk

Licia Capra
University College London e-mail: l.capra@cs.ucl.ac.uk

objects, VGI has improved the state of geographical information systems greatly over the past years by engaging participants from all the world. OpenStreetMap (OSM) alongside WikiMapia is perhaps one of the most successful examples of VGI, with over 1.7 million users [27] collectively building a free, openly accessible, editable map of the world. However, these platforms have been subject to scrutiny by the research community over the years; one of the fundamental concerns is the credibility and integrity of the contributed information, as we take a task away from skilled employees and assign it to an undefined, self-selected crowd. Several studies have investigated the credibility of the volunteered content [7, 9, 11, 14, 19, 28, 30] and have concluded that the content is of high quality. However, the quality of the contributed information is not the only concern that has emerged as a result of this knowledge production paradigm shift; another important issue is that of completeness of the information [11, 15]. That is, how much information is contributed to the VGI systems and how would such a system grow over time. In addition to these two concerns, researchers have also been addressing the voluntary dimension of VGI by looking at what incentive models and factors motivate the contributions from individuals [4, 13]. However, one aspect that has perhaps received less attention from the GIS research community is the impact of *society* and factors such as those of socio-economics on the contributed information. This aspect has been studied extensively in the domain of social sciences [12, 24] and has been shown to have a high impact on the content generation in Web 2.0. In the domain of GIS, as the contributions are intrinsically spatial, the potential digital production gap may contribute to some areas not being mapped. The risk of course is that if the socio-economic factors are responsible for this digital production gap, the deprived areas (*e.g.*, less wealthy) would also remain information deprived [8]. We thus aim to investigate this aspect further by studying the extent to which society's characteristics such as *socio-economic* factors determine the success of spatial crowdsourcing and VGI.

To this end, in this chapter, we study the impact of society on the quality and completeness of the contributed information in OSM. In particular we investigate over 6 years of OSM contributions to Greater London, United Kingdom, in terms of its accuracy and completeness. We first measure the quality of OSM contributions in terms of positional accuracy and thematic accuracy, paying particular attention to what properties of the editors are responsible for this quality. After presenting the accuracy of OSM in London, we then investigate the impact of society in the completeness of this contributed information. We do so by first measuring the completeness by comparing the OSM data to a proprietorial dataset. We then argue that this completeness is affected by the socio-economic factors of the society and propose a model that can capture the completeness as well as its evolution over time.

We finally end this chapter by presenting the possible interventions that could be done to improve communities' participation in OSM. We do so by measuring the impact of the social engagement of the editors on their participation, which can then in turn account for improving the sparsity of the contributed information.

2 OpenStreetMap London

OpenStreetMap is freely available to download from various repositories on the web which provide the latest snapshot of the OpenStreetMap project. We gathered the dataset of London, United Kingdom, from [6] which contains the history of all edits since 2006 on all spatial objects performed by all users. In OSM terminology, spatial objects can be one of three types: *nodes*, *ways*, and *relations*. Nodes are single geospatial points, defined using latitude/longitude coordinates, and they can be used to represent Points Of Interest (*e.g.*, cafes, restaurants, hospitals, schools); 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).

We chose Greater London, United Kingdom as our subject city as we are interested in studying a metropolitan city with a diverse society which has been engaged with OSM since the very beginning. Furthermore, as we are interested in evaluating the impact of society on the contributed information, we limited our investigation to only Points Of Interest (POIs) rather than the roads. This is because the contribution to the OSM differ greatly between the two categories: road mapping is typically done by users who have high expertise in both the geography of an area and the editing tools required to digitally represent it, while POI mapping can be performed by any city dweller, with no specific cartographic skills required. The latter category is thus more representative of the broad VGI setting and the impact that ordinary citizens can have on VGI. Finally, to consider only genuine users' contributions, we have excluded contributions that most likely correspond to bulk imports. Two bulk imports were detected in the whole dataset, with tens of thousands of edits done in a single day by a single user, spread throughout Greater London (*e.g.*, more than 20,000 post boxes spread across all Greater London appeared in OSM in only one day in 2009 from the same user). We chose to discard such data as we intend to model genuine 'bottom-up' user-generated contributions, of which massive imports are not representative of.

To evaluate the quality of the OSM POIs in London based on the dataset at hand, we need to: (1) define benchmarks against which to compare accuracy; and (2) define quality metrics for OSM objects.

Benchmarks

We considered two different commercial geographic information systems covering the same type of information (in terms of POIs) as OSM: *Navteq* and *Yelp*. Navteq [21] is 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. Yelp [29] focuses on business listings, from store-fronts (*e.g.*, restaurants and shops) to services (*e.g.*, doctors, hotels, and cultural venues). Being

commercial services, Yelp and 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). We then built our benchmark (or *ground truth* dataset) as the set-intersection of Navteq and Yelp data; in doing so, a POI in Navteq is considered to be the same POI in Yelp if the name is the same and the geographic distance is less than 20 meters.

Metrics

In both OSM and in the ground-truth dataset, a POI is defined as a triple: $poi = \langle name, amenity, (lat, lon) \rangle$, where *name* is the POI's name, *amenity* is its category (*e.g.*, cafe, restaurant), and (lat, lon) are the coordinates defining its geographical position. We then quantify *quality* of OSM data in terms of its positional accuracy and thematic accuracy defined based on *geographic error* and *lexicographic error*, respectively. We measure geographic error as the Euclidean distance between the OSM points and those in the ground truth dataset. The lexicographical error is computed as the Levenshtein distance between the POI names (of OSM and the ground truth dataset). This calculated Levenshtein distance captures the minimum number of single characters that are required to change the POI name as stated in OSM to the name that exists in the ground truth dataset. Finally, we consider two points to be equivalent in both datasets if their geographical error is less than 100 meters and their Levenshtein distance is less than 0.35[1].

The results based on the above benchmark and metrics indicate an overall *high quality* of information for OSM POIs with geographic errors almost normally distributed and their average value is less than 25 meters thus revealing accurate positioning of POIs on the map with respect to the ground truth dataset. Lexicographic error is almost zero (0.13 on average), thus revealing thematic accuracy in spelling names of POIs. This overall high level of quality in OSM is an extremely positive result, encouraging further insights to measure the completeness of this information.

3 Impact of Society on Information Completeness

In the previous section we demonstrated that the OSM POIs have a high quality. Indeed, OSM's positional and thematic accuracy has shown to sometimes supersede the most reputable geographic datasets, performing especially well in urban areas [11]. However, these accuracy measures alone are not enough, and another concern is the *completeness*. In other words, *what part* of the physical world has been digitally mapped and which parts are lacking digital representation? Answering this question is necessary, in order to understand *where* crowd-sourced map information

[1] These values were chosen after manual inspection of a number of POIs jointly present in the two datasets that we knew to be the same.

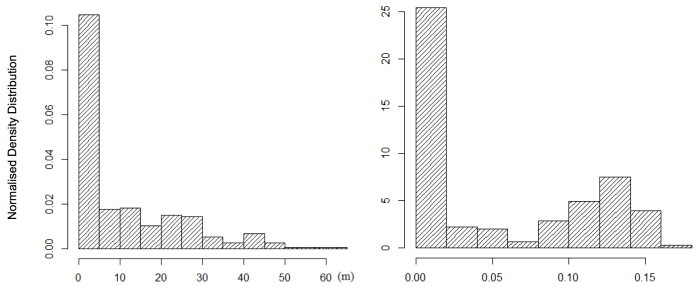


Fig. 1 Normalized density distributions of standard deviation of error for geographical error and lexical error, respectively.

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. To address this question, we investigate the impact of society on the completeness of information. In so doing, we first measure the completeness of information in terms of POI presence in OSM London in comparison with the described benchmark dataset.

In order to compute the completeness of POIs, we require a matching algorithm to map OSM POIs to those in the benchmark dataset. We first need to relate POIs in OSM with the same POIs in the ground-truth dataset in an automatic way. We borrowed the same matching methodology as described earlier, where two POIs are considered the same based on their lexical and geographical similarity.

Based on the above matching, we have evaluated completeness of OSM POIs for Greater London as:

$$\text{Completeness} = \frac{\#(\{\text{POIs in OSM}\} \cap \{\text{POIs in Ground} - \text{Truth}\})}{\#\{\text{POIs in Ground} - \text{Truth}\}} \tag{1}$$

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

A Non-uniform Completeness

This section reports on the results of our completeness analysis based on the above formulation. We first considered the area of Greater London as a whole, for which we found the completeness to be 0.35. However, this single aggregate value does not reveal much in terms of what areas of London are being digitally mapped. We thus considered the finest level of granularity for which societal information (such as population *etc*.) is still available. We selected *wards* representation to define the spatial granularity of our analysis. Wards are spatial boundaries defined by London Local Authorities. Currently London consists of 600 wards [18]. Figure 2 illustrates the choropleth map of London's POI distribution, where each tile represents a ward.

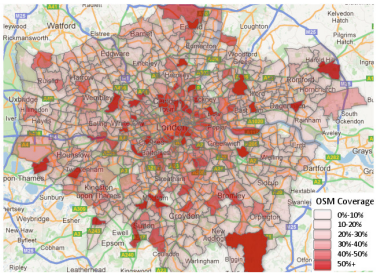


Fig. 2 Choropleth map of OSM POI's completeness for Greater London. The darker area correspond to higher completeness.

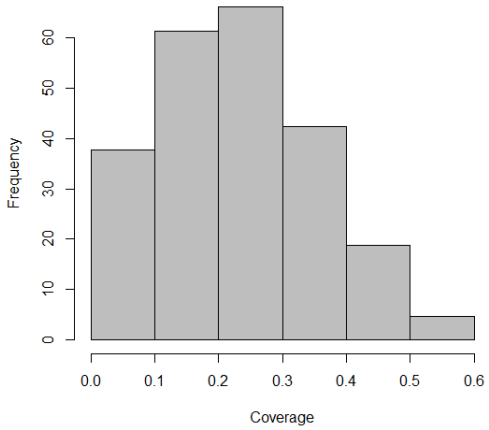


Fig. 3 Frequency distribution of POI completeness.

As shown, completeness is non-uniformly distributed across the city. Previous studies on completeness of OSM for road networks have revealed that distance from the city center is inversely related to this completeness [30]; although at a first approximation a similar pattern seems to emerge for POIs too (i.e., the further away we move from the city center, the worse the completeness), we can also identify various suburban areas with high completeness. Figure 3 further shows the histogram approximating completeness distribution at ward level. As shown, there are many wards where completeness is very low (≈ 0), and a few wards where completeness is quite high (≈ 0.6) instead.

We hypothesize that the society's characteristics influence and contribute to this non-uniform distribution of spatial information. To test our hypothesis we take a closer look at the contextual factors affecting different areas (wards) of London. We extract and consider the following society factors:

Population. Using UK Census 2011 data published by the National Statistics Office [1] we have information about population at the 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 [7]. 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 a higher number of contributors, we may expect more contributors per unit area to exist in denser areas. The hypothesis we thus want to test is the higher the *population density* of an area (that is, population divided by ward size), the higher the completeness.

Poverty. Analyzing the relationship between poverty of an area and completeness is important, as it may reveal the impact that (lack of) technology adoption (*e.g.*, use of 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 is 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. The hypothesis under test is that poverty of an area is negatively correlated with digital mapping of its POIs.

In addition to that we consider in our study the distance from where the social and economic activities happen. Previous studies on OSM have shown that road completeness decreases when moving away from the city centers [30]. Similarly, we are interested in examining the effect of distance from the city center on completeness. However, metropolitan cities contain more than one center *per se* but include multiple urban hubs referred to as poly-centers [2]. London currently has 10 different poly-center [23], from which we consider and compute the Euclidean distance. We then used the shortest distance as our 'distance from the center' factor, and tested the hypothesis that the closer to the center, the higher the completeness.

Table 1 reports the Pearson Correlation coefficients between each of the previous factors and OSM completeness as well as the p-value codes, indicating the significance level of each presented result.

Factor	r	*p-value*
Population Density	0.32	***
Poverty	-0.10	*
Distance from the Nearest Poly-center	0.36	***

Table 1 Pearson correlation coefficients r and p-values codes between socio-economic factors and OSM completeness at the ward level (p-value significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1).

The results indeed confirm that population density is positively correlated with completeness ($r = 0.32$ and p-value < 0.001). In particular, we found that an increment in population density of 50 people per hectare corresponds to a 25% increase in completeness for the average case. Focusing on poverty, we can see that $r =$ -0.10 is significantly weaker than that found for other factors such as population, suggesting that, although significant, poverty itself is only a secondary factor in explaining completeness. Turning our attention to the last factor under examination, distance to the closest poly-center, our intuition is confirmed by Table 1, which shows that distance

from the closest poly-center is inversely correlated with completeness ($r = 0.36$ and p-value < 0.001). In particular, we also found that a decrement of 5 km in distance from the closest poly-center corresponds to a 28% increase in completeness for the average case.

However, these findings raise concerns in terms of the long-term sustainability and completeness of the VGI. More specifically, is the completeness going to spontaneously grow across the city? Or are there going to be areas that will continue to be neglected? To address this question we built a set of models based on the discovered socio-economic factors that can accurately capture the *digital* growth of spatial information in VGI.

Growth of Spatial Content Production

In order to measure the growth of contributed information over time, the first step was to choose a *spatial* and *temporal* unit of analysis. In terms of the spatial unit of analysis, we have maintained the same level of granularity as before and operate at the ward level of London. In terms of the temporal unit of analysis, we tried different time units, from finer (3 months) to coarser (18 months) granularity. In the end, we chose to report the results for the smallest unit of granularity (12 months) that still afforded statistically significant results across *all areas* of Greater London.

We then needed to define a metric that reflected which areas had been digitally mapped and which had been neglected instead. To this purpose, it is worth pointing out that not all areas naturally require the same amount of OSM edits to be mapped. For example, areas containing many services and attractions will require many OSM edits to be mapped (*e.g.*, Soho in London); however, sparse areas like parks and industrial estates will require significantly less. To capture this property, we chose as metric *OSM activity*, defined as the number of OSM edits *relative to* the number of physical POIs in each ward at that time:

$$\text{OSM activity} = \frac{\#\text{OSM edits}}{\#\text{POIs}} \tag{2}$$

#OSM edits is readily available from our OSM dataset. To estimate #POIs, that is, the actual number of POIs present in each area, we used the ground-truth dataset as before.

Figure 4 illustrates the cumulative temporal evolution of OSM activity (Equation 2) in London from 2007 to 2012. As shown, the vast majority of areas have low cumulative activity (with only a few wards slightly above 0.5); furthermore, complex dynamics are at play, with no clear pattern emerging (*e.g.*, no core-to-periphery spreading).

To capture the growth of OSM information, we built a regression model which takes into account the past OSM activity of each area as well as the following two features:

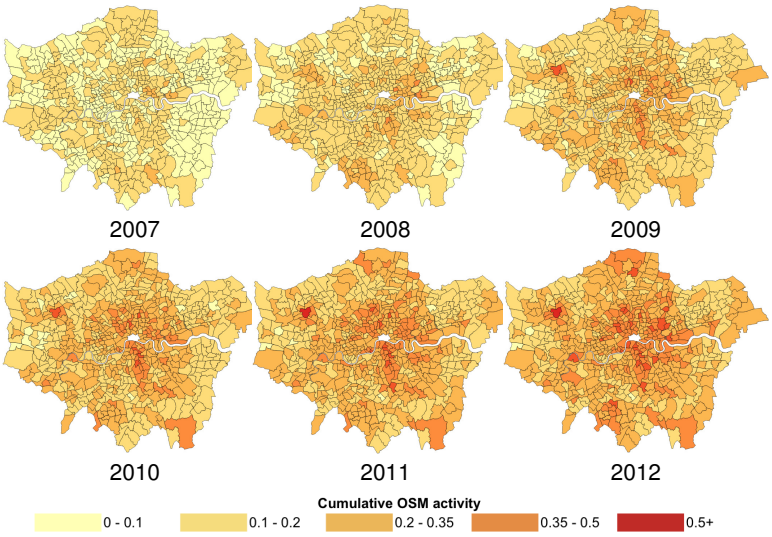


Fig. 4 Cumulative OSM activity from 2007 until 2012.

Community Editing. We argue that, regardless of spatial positioning of wards within a city, they attract the same OSM contributors (because they might, for example, offer related attractions/urban functions [5]). We thus incorporate the *community* feature into our model which hypothesizes that if a ward has been edited by contributors who have heavily edited other wards in the past year, the ward is likely to be edited in the future. In other words, the activity of a ward depends on the past activity of its 'co-edited' areas, where two areas are defined as 'co-edited' if they are edited by the same shared community of editors.

Society Factors. Based on our hypothesis that the society has an impact on the VGI, we incorporate population, poverty and distance from the center into our model. We hypothesize that these societal factors influence the likelihood of a ward being mapped in the future.

We then built a linear regression model where the predicted outcome of OSM $act(w_i, t+1)$ - the activity in a ward w_i at time $t+1$ based on the above features.[2]

To quantify the *predictive* accuracy of our model we then conducted a classification experiment, and used the discovered classification parameters to classify OSM activity for the upcoming year. For example, we used 2007/08 to estimate the parameters, built our model, then made predictions for 2009. In this case, we divided the outcome of our models into two distinct categories: 'slow future OSM activity growth' (when $act(w_i, t+1) < 0.3$) and 'fast future OSM activity growth' (when $act(w_i, t+1) \geq 0.3$), with 0.3 being the median value of OSM activity growth for the time windows under consideration. Finally, we considered the top 75% wards in London only, as predicting OSM activity growth of very sparse areas (*e.g.*, parks)

[2] The full details about the model can be found in [22].

Predicted Year	*TN Rate*	*TP Rate*	*Accuracy*	*Senitivity*
2009	0.78 (+56%)	0.75 (+50%)	0.77 (+54%)	0.77 (+54%)
2010	0.81 (+62%)	0.78 (+56%)	0.80 (+60%)	0.80 (+60%)
2011	0.82 (+64%)	0.79 (+58%)	0.81 (+62%)	0.81 (+62%)
2012	0.79 (+58%)	0.75 (+50%)	0.77 (+54%)	0.78 (+56%)

Table 2 True Negative Rate (slow growth), True Positive Rate (fast growth), Accuracy and Sensitivity of our classification model. Relative improvement of each model with regards to a random classifier is also reported in parentheses.

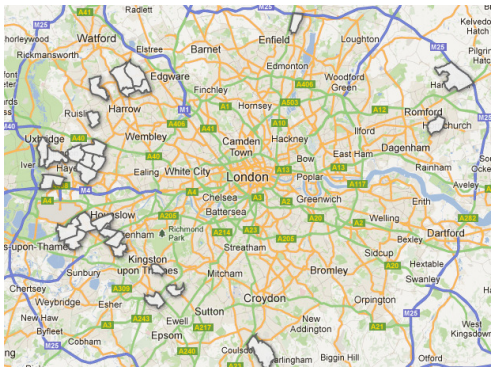


Fig. 5 OSM in 2013. Highlighted are the wards for which our model predicated a slow growth.

has little significance. Table 2 presents the results of the classification. As shown, the accuracy of our model is quite high with up to 82% for slow growth and 79% for fast growth.

Being able to *predict* what areas will not be digitally mapped can help to plan and execute interventions. Such interventions may span a wide spectrum: from allocating financial resources to cover neglected areas, to organizing public mapping events to direct the crowd towards specific mapping goals. Having an accurate growth model at hand implies that these limited resources (human and/or financial) can be best allocated to maximize return on investment. For example, Figure 5 illustrates the wards of London for which our classification model forecasts slow OSM growth in 2013; various wards in the west/north-west/south-west of London are highlighted as areas at risk. Using influence maximization schemes (*e.g.*, [25]), one could decide how many resources to allocate to each of these highlighted areas, to maximize expected growth in the following year(s), both as an immediate result of investment *and* thanks to the contagion and self-reinforcement processes that should follow.

4 Impact of Social Mapping Parties

So far we have looked at the impact of society and socio-economic features on the contributed information in OSM in terms of spatial completeness and information growth. We now focus on what motivates people to contribute to OSM, paying particular attention to the social aspects of the OSM community. Social contact

has been identified as a powerful motivator by many successful online communities [16], Hackathons, mapathons and other similar social events are often organized, in order to bring together people with similar technical skills and interests to accomplish collaborative projects. Likewise, OSM contributors organize local social events, so called mapping parties throughout the year, to bring together the editors to *socialize*, *map*, and *engage the new comers*. In London, these events happen on a fortnightly basis [10] and their details (when/where it happened as well as who participated) are recorded on OSM wiki pages [26].

To understand whether these mapping parties are successful in encouraging participation, we address the following two research questions: i) Do the mapping parties cause users to map more than usual *during* the collaborative event? ii) Do the mapping parties cause users to map more than usual afterwards both in the *short* and *long term*?

To address these questions, we borrow from the field of economics [20] and quantify the direct (immediate) and indirect (subsequent) impact of a mapping party using the Abnormal Returns (AR) model. ARs are triggered by events, in our case the mapping party, and are assessed as the higher the abnormal return, the higher the impact of the event (mapping party) on the variable (user contributions in our case). In our analysis we define for each user i and time period τ after the party, we measured the *actual* returns R_i^τ as the average number of contributions per unit of time Δt made by user i during period τ. We also computed the *expected* returns E_i^δ as the average number of contributions made by the same user i per unit of time Δt during a period δ prior to the event. We then calculated the abnormal returns $AR_i^{\delta\tau}$ per unit of time Δt of each user i as:

$$AR_i^{\delta\tau} = R_i^\tau - E_i^\delta \tag{3}$$

In order to conduct impact analysis for mapping parties on users' contributions, we needed to manually construct dataset of the mapping parties in London both in terms of where it happened (the geographical area) and who took part in the event. We recorded 94 mapping parties for the period under examination. As we do not have ground truth about who took part in what event, we inferred a set of 150 'social mappers' from the list of users in the wiki who 'intended to attend' and had made an edit during the event time in the vicinity of the mapping event. Figure 6 illustrates an example of this inference for a mapping party that took place in the Isle of Dogs area of London.

Direct Impact of Mapping Parties

The first hypothesis we tested is that users contribute more during mapping parties than outside these events. For each mapping party, and for each user who took part in it, we compute the abnormal returns as per Equation 3, with Δt equal to one day. We further selected δ equal to six months prior to each party, to have enough history

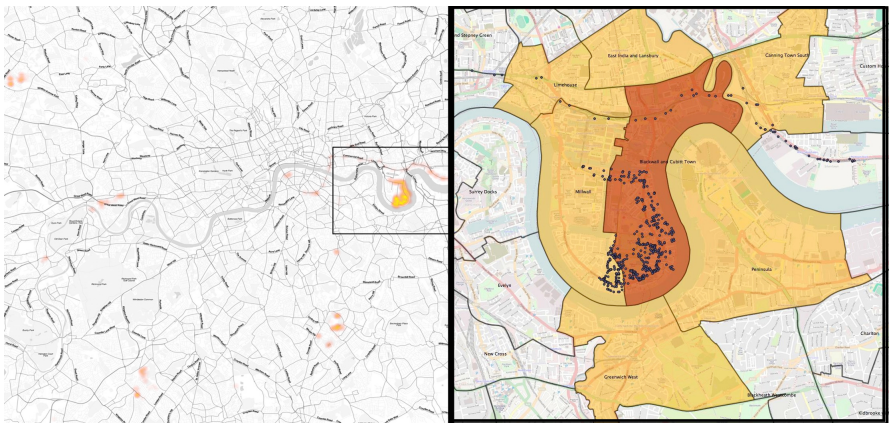


Fig. 6 Map of edits made around London over 48 hours during and after one of the identified mapping parties (the Isle of Dogs mapping party).

about users' editing behaviour, and τ equal to the 'party time' (from the day of the party up until midnight of the day after).

OSM users greatly differ in terms of the amount of contributions they make, and over what timespan [14]. In order to quantify the impact of mapping parties on different types of users, we have grouped them based on the number of contributions they made in the six months prior to each party. We do so on a log scale of 10 as in the above pre analysis, and split users into five distinct groups: *Group 0* (just 1 edit); *Group 1* (from 1 up to 10 edits); *Group 2* (from 10 up to 10^2 edits); *Group 3* (from 10^2 up to 10^3 edits); *Group 4* (from 10^3 up to 10^4 edits). An additional group of newly joined users (*Group NA*) is considered, consisting of those who make their first edit in the system either during the mapping party or less than six months preceding it (thus not having sufficient editing history to be confidently placed in the above groups). The results for this group assess the impact of mapping parties on new comers.

Figure 7 shows average results across the 94 mapping parties that took place in London in the period under study, for each of these user groups. We use a box-and-whiskers plot, with the thick black line within each box representing the median value and the 'whiskers' of the box representing the top and bottom quartile values. Median y values above zero indicate that most users within that group exhibit a higher number of edits during the party time than before it, and vice versa (negative y values indicate reduced activity during the mapping as compared to the norm).

The results show that for Groups 0–2 (light to medium contributors) and Group NA (new comers), mapping parties have a strong positive impact in terms of contributions, with their edits being significantly more than usual. Despite more variation within it (and some negative returns too), Group 3 experienced the overall highest AR, with more than 50% of its members (median value and above) contributing at least 100 edits *more than expected* in the observation period (i.e., party time). Perhaps surprisingly at first glance, only half of the heaviest editors (Group 4) con-

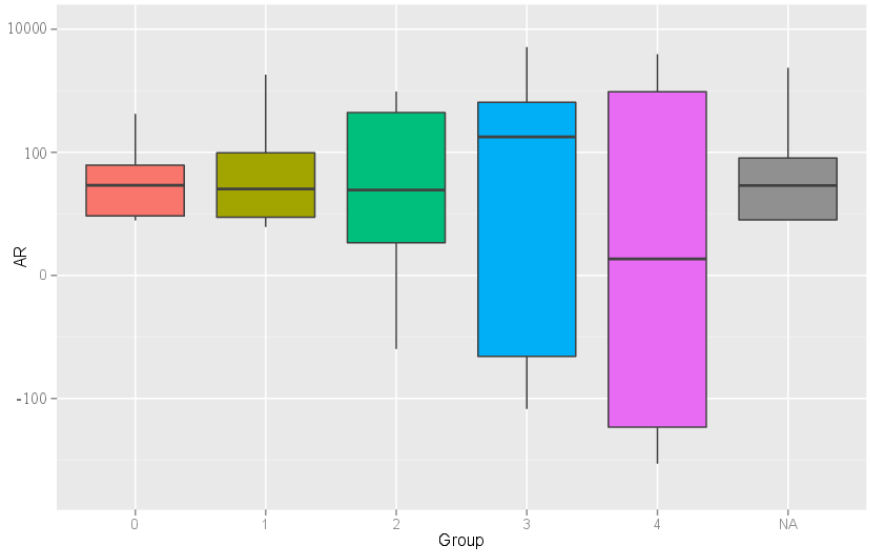
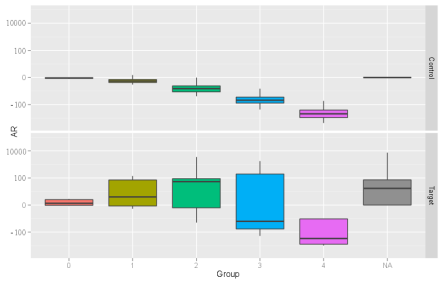


Fig. 7 Box-and-whiskers plot of abnormal returns during a party.

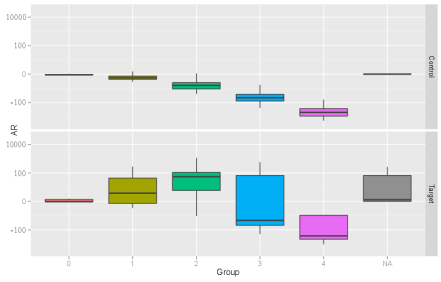
tribute more than expected; the other half in fact perform much below par. We cross checked the names of some of these contributors against what is publicly available in OSM wikis, and found that many of these users take on organizational roles, visiting an area prior to the party, creating 'cake diagrams', and identifying 'problems' they would like the party to fix. We thus speculate that their reduced contribution during the event itself might be due to their engagement in organizational rather than editing activities (*e.g.*, acting as demonstrators for less expert users).

Indirect Impact of Mapping Parties

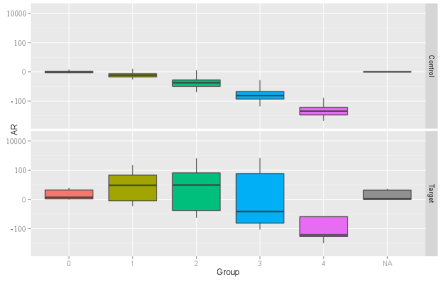
The second hypothesis aims to quantify the impact that mapping parties have on users' contributions *after* they took part in an event. As before, we do so by computing AR for the 6 user categories (from light to heavy editors: Groups 0–4, and new comers: Group NA). To distinguish between the impact caused by attending a party from the impact potentially caused by external events (*e.g.*, weather, OSM advertising), we constructed control groups for each of the six study groups. Each respective control group includes users who (i) have had a similar number of contributions as users in the corresponding study group in the $\delta = 6$ months prior to the party under examination and (ii) who did not take part in it or any other event in that time period. We then computed AR for each control group too.



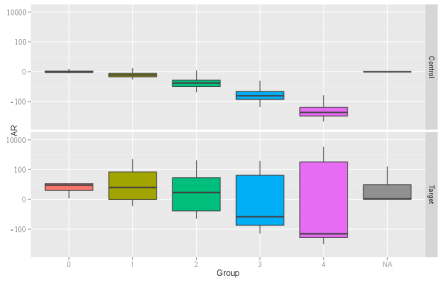
a. 1 week following a mapping party.



b. 1 week to 1 month following a mapping party.



c. 1 month to 3 months following a mapping party.



d. 3 to 6 months following a mapping party.

Fig. 8 Box-and-whisker plots of abnormal returns for the mapping parties.

To quantify both short- and long-term effects of mapping party attendance, we computed AR on four non-overlapping observation windows τ: (i) up to one week following the event, (ii) between one week and one month following the event, (iii) between one and three months following the event, and (iv) between three and six months following the event. All observations exclude the contributions made *during* the event. For an easy comparison across all plots, we chose Δt equal to one week as the unit of time to compute AR across all cases. Results for each observation window are shown in Figure 8. Once again, we use box-and-whiskers plots, with boxes in the upper part of the plot illustrating the behaviour of the control groups, and the bottom part displaying the behaviour of the study groups (referred to as 'Target' group in plots).

First of all, we observe a decline in contributions (negative AR) by all *control* groups across all observation windows: users who do not take part in a mapping party tend to become more and more disengaged as time passes. This loss of engagement is more pronounced for users who were previously heavily contributing to OSM (Groups 3 and 4).

Conversely, both light contributors and medium contributors (Groups 1 and 2) who attended a mapping party tend to be more engaged over time (in both 1-to-3 and 3-to-6 months – see Figures 8 c and d). As for the heavy contributors (Groups 3 and 4), we observe slightly increased engagement in the short term (Figure 8). However, as time progresses (Figure 8 b), we observe 25% of Group 4 participants now exhibiting positive abnormal returns, whilst the AR of its control group remains consistently low. Finally, in the longer term (3–6 months, Figure 8 d), Group 4 is indeed the only study group exhibiting significantly more engagement than what is observed in the corresponding control group. As for Newcomers (Group NA), we can observe that a strong positive AR is indeed evident in the first week following participation in a mapping event (Figure 8). However, after the first week following the event, 50% of the newcomers stop contributing completely, with further complete disengagement as time passes.

5 Summary and Conclusion

In this book chapter, we studied the impact of society on the crowd-sourced spatial information for OSM London. We showed that the positional and thematic accuracy is high, while the completeness of this information is low and non-uniformly distributed across the city. We revealed that different societal factors, including population density, distance from the center and poverty, are correlated with the information completeness. Given the role that these factors play on the production of digital spatial content, the risk that arises is that the deprived areas might also remain *digitally deprived* on the maps. As a result they may attract even less attention from visitors and city dwellers, thus putting their economy at risk [3, 17].

However, as we presented this low completeness is not a problem *per se*, if we are able to model the information growth and digital production based on societal

factors. Indeed, being able to build a model that *explains* growth, and that accurately detects what areas are most likely to suffer from neglect, has enabled us to highlight these areas and so to bring them to the attention for targeted interventions. One form of these targeted interventions is the design of incentives to call the OSM community to edit specific areas by participating in OSM mapping parties. In understanding whether such an incentive model is successful, we studied OSM mapping parties and revealed that they are extremely successful in retaining users and increasing the editors' contributions both in the long and short term. However, the demonstrated results are based on the users who attended the pre-organized London mapping parties. Therefore, further research is required to understand how communities react to the directed mapping parties, and how the success of the mapping parties may vary across different culture traits around the world.

References

1. Census 2011. http://www.ons.gov.uk/ons/guide-method/census/2011/.
2. S.D. Brunn, J.F. Williams, and D.J. Zeigler. *Cities Of The World: World Regional Urban Development*. Rowman & Littlefield Publishers, 2003.
3. Susan Cain. *Quiet: The power of introverts in a world that can't stop talking*. Random House LLC, 2013.
4. David J Coleman, Yola Georgiadou, Jeff Labonte, et al. Volunteered geographic information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, 4(1):332–358, 2009.
5. J. Cranshaw, R. Schwartz, J.I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.
6. Geofabrik. http://www.geofabrik.de/data/download.html.
7. J.F. Girres and G. Touya. Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14(4):435–459, 2010.
8. Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, (ahead-of-print):1–19, 2014.
9. 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.
10. M Haklay and P . Weber. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.
11. M.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. *The Cartographic Journal*, 47(4):315–322, 2010.
12. 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.
13. Francis Harvey. To volunteer or to contribute locational information? towards truth in labeling for crowdsourced geographic information. In *Crowdsourcing Geographic Knowledge*, pages 31–42. Springer, 2013.
14. J Jokar Arsanjani, C Barron, M Bakillah, and M Helbich. Assessing the quality of openstreetmap contributors together with their contributions. In *Proceedings of the 16th AGILE conference, Leuven, Belgium*, 2013.

15. Thomas Koukoletsos, Mordechai Haklay, and Claire Ellul. Assessing data completeness of vgi through an automated matching procedure for linear data. *Transactions in GIS*, 16(4):477–498, 2012.
16. R. Kraut and P. Resnick. *Building Successful Online Communities: Evidence-Based Social Design*, pages 42–43. The MIT Press, 2012.
17. Stephen EG Lea. *The individual in the economy: A textbook of economic psychology*. CUP Archive, 1987.
18. London Data Store. http://data.london.gov.uk/datastore/package/ward-profiles-2011
19. 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.
20. A. Craig MacKinlay. Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39, 1997.
21. Navteq. http://www.navteq.com/.
22. Giovanni Quattrone, Afra Mashhadi, Daniele Quercia, Chris Smith-Clarke, and Licia Capra. Modelling growth of urban crowd-sourced information. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 563–572. ACM, 2014.
23. Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthlemy. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1), 01 2011.
24. J. Schradie. The digital production gap: The digital divide and web 2.0 collide. *Poetics*, 39(2):145–168, 2011.
25. Yaron Singer. How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *Proc.of the 5th ACM International Conference on Web Search and Data Mining*, pages 733–742, 2012.
26. OSM Wiki. http://wiki.openstreetmap.org/wiki/london/summer_2008_mapping_party_marathon/2008-05-21.
27. OSM Wiki. http://wiki.openstreetmap.org/wiki/stats.
28. D.M. Wilkinson and B.A. Huberman. Assessing the value of cooperation in wikipedia. *First Monday*, 12(4), 2007.
29. Yelp. http://www.yelp.com/.
30. 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.