GeoJournal (2018) 83:427–440
https://doi.org/10.1007/s10708-017-9778-7

Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks?

Jonathan Bright · Stefano De Sabbata · Sumin Lee

Published online: 25 May 2017

Abstract Crowdsourced knowledge websites such as Wikipedia and OpenStreetMap are increasingly attracting a critical literature which has highlighted the fact that the contributor bases of these sites are often geodemographically biased: drawn from more affluent and better educated segments of the population. However, while bias in contributors is well known, we know less about whether this also results in a bias in outcomes on these websites: or whether the partial portion of the population which does make contributions also works to "fill in the blanks", by adding knowledge about other less well-off neighbouring areas which have not attracted a contributor base. This article addresses the question of whether such "neighbourhood effects" exist in practice. It makes use of a novel dataset of alcohol license data in the UK to assess variation in the completeness of the volunteer geographic information site OpenStreetMap. The results support existing literature in showing that completeness is related to demographics: areas with higher levels of wealth and education typically exhibit higher levels of completeness. The article then makes a novel contribution by showing evidence of the existence of neighbourhood effects: poorer regions with more affluent neighbours typically having higher levels of completeness than poorer regions which are also surrounded by poorer neighbours. The results suggest that crowdsourced knowledge websites can aspire to a kind of completeness even whilst user bases remain partial and biased.

Keywords OpenStreetMap · Crowdsourcing · Volunteer geographic information · Participation inequality

J. Bright (✉) · S. Lee
Oxford Internet Institute, University of Oxford, Oxford, UK
e-mail: jonathan.bright@oii.ox.ac.uk

S. De Sabbata
University of Leicester, Leicester, UK

Introduction

Crowdsourced knowledge websites, which encourage users to contribute to a collective endeavor of knowledge creation, are having a profound effect on how information is both produced and distributed within societies. Wikipedia, an online encyclopedia created by the volunteer contributions of thousands of editors, is only the most well-known example of this phenomenon, which also includes the mapping website OpenStreetMap, citizen science platforms such as Galaxy Zoo, thousands of expert question and answer forums such as those run by StackExchange, and many more. These sites vary in aim and design, yet their guiding principle remains the same: people volunteer to create the information on the site, usually without the prospect of direct personal benefit, and this information is then freely released to the world at large. The data created is being put to use in a wide

variety of academic, government and business contexts (see e.g. Voigt and Bright 2016).

Despite the obvious normative appeal of freely created and distributed information, recently critical appraisals of crowdsourced knowledge have also started to appear. As well as those questioning the accuracy of data created by those who do not (necessarily) have formal qualifications or expertise (Senaratne et al. 2017), a strong line of critique has been the potential emergence of geodemographic "biases" within the data. The groups of people creating this information, it has been noted (see e.g. Budhathoki and Haythornthwaite 2013), are typically drawn from more affluent and educated parts of the population at large: and this might be expected to create, naturally, a slant in the types of topics and issues they are interested in. Hence volunteer information as a whole may be biased towards the views of those creating it.

However, despite the importance of these concerns, research in this area is still in its early stages, and hence we have only partial information about the extent to which geodemographic biases truly make a difference in the outcomes of crowdsourced knowledge websites. In particular, although it is quite clear that the contributor base is biased, we do not know if bias in contributors always results in a bias in outcomes; or if a geodemographically unrepresentative group of contributors can nevertheless work to "fill in the blanks" by providing information which might be outside of their own personal experiences or interests. For example, research has demonstrated that while Wikipedia suffers from a dearth of contributors from Africa, it also has many articles on African subjects which are contributed by people living elsewhere (Graham et al. 2014). Hence one could argue that a lack of information production in one area is ameliorated by information production in other neighbouring areas (of course, whether this replacement is perfect, or even desirable, is an important secondary question, to which we will return in the conclusion). However, as yet no research has tried to address systematically how widespread these neighbourhood replacement effects are, or quantify their importance.

This paper aims to address this deficit. The research question is simple: to what extent do "neighbourhood effects" moderate geodemographic biases in crowdsourced knowledge websites? By neighbourhood effects, we mean the existence of a dynamic whereby information relating to an area with few participants is created by neighbouring areas which have more participants, thus providing a kind of compensation for a lack of contributions from this area.

The potential existence of neighbourhood effects is of crucial practical significance. If they do exist, it would suggest that crowdsourced knowledge projects can aspire to something like completeness even if the demography of their contributors is highly biased. Indeed, it would suggest that recruitment strategies for these websites should simply focus on maximizing the volume of participants (rather than their diversity). By contrast, if they do not exist, then such projects will always be destined to produce only a partial version of reality while their contributor bases itself remain biased. This would suggest the need for recruitment strategies to focus on attracting new types of people to the project.

We address our question by focusing on data present in OpenStreetMap [OSM] in the UK. OSM is a good case study for this type of question because the data within it can easily be attached to a given geographic area, for which census level demographic data is available. Making use of a unique dataset of alcohol licenses granted in the UK, we assess in particular the completeness of OSM "point of interest" data, which relates to geographic features which are located at a single point such as shops, schools, hospitals and restaurants. Unlike so called "linear" features such as roads, the fact that points of interest can be located at a single point (or at least a small area) means that we can directly address questions about the impact of local area demographics on their completeness, and also the potential impact of neighbourhood effects.

Our results show that OSM is just under 50% complete in terms of the presence of our specific alcohol license data; however, there is reasonably high geographic variation in this completeness. Building on existing literature, we first test a range of demographic hypotheses to try and explain this variation, showing that the average education level and average wealth of an area are correlated with higher completeness, whilst more densely populated areas typically had lower license completeness. We then address our key question of neighborhood effects, and show evidence that better educated and economically well off areas do indeed appear to compensate for the lack of OSM development in their neighbours: indeed, worse off regions which have well off neighbours tend to have similar levels of completeness as regions which are

well off themselves. We discuss these results in terms of a continuing digital divide within crowdsourced knowledge websites and hence build theory about the future scope of these projects.

Completeness in crowdsourced knowledge websites

In this first section, we will review existing literature on demographic biases in "crowdsourced knowledge websites". We define these websites as locations on the internet where groups of people contribute information to a collective endeavor of knowledge building; knowledge which is then made available for free to the world at large. Crowdsourced knowledge is a phenomenon which can be found in a wide variety of fields and as such has attracted a variety of more domain specific labels. Wikipedia, probably the most famous example of such a website, is often defined in terms of "crowdsourcing"(Shirky 2008), which is the term we use in this article. However, other terms are widely used (See et al. 2016). For example, freely contributed news stories and journalism have been labelled "user generated content" or "citizen journalism" (Goode 2009; Jönsson and Örnebring 2011); contributions to volunteer mapping initiatives have been called "volunteer geographic information" (Goodchild 2007); and participation in academic research has been labelled "citizen science"(Bonney et al. 2014). Although divergent in subject matter, all of these terms try and address the same fundamental activity: time freely given to a "knowledge building community" (Rafaeli and Ariel 2008).

As the popularity and impact of these sites has grown, so too has academic research on the subject. Initially, this research was largely focused on the accuracy of data, which is unsurprising given the lack of formal qualifications of participants in crowdsourced knowledge sites, or at least formal vetting processes for establishing these qualifications (Senaratne et al. 2017). Much of this work has been cautiously optimistic, noting that data is incomplete yet nevertheless of a reasonable standard, and hence can be made use of if people are conscious of potential weaknesses in the data.

A second area of research has been the nature of participants in these sites. Crowdsourced knowledge projects are typically characterized by openness: anyone can in principle become a contributor, and indeed most such projects encourage anyone who might be interested to sign up (Franzoni and Sauermann 2014). In addition to being free to join, such efforts are of course also free to leave: meaning that many people who do contribute do so for only a short amount of time, whilst the majority of content is produced by a small minority of committed users (Cobo et al. 2016; Haklay 2016; Panciera et al. 2009; Sauermann and Franzoni 2015). Hence the nature of the people that do end up both joining and in particular staying is a key area of research interest.

One of the most important, and consistent, findings from research in this area has been to show that the demographic makeup of participants in crowdsourced knowledge projects is "biased": i.e., it often differs considerably from the makeup of society as a whole. Perhaps the most striking finding concerns gender, with studies repeatedly documenting that contributors to crowdsourced knowledge sites are overwhelmingly male. For example, females constitute less than 30% of the editors of Wikipedia (Hill et al. 2013). However demographic inequality has also been documented in the areas of wealth and education (see e.g. Budhathoki and Haythornthwaite 2013; Glasze and Perkins 2015; Haklay 2010), with participants typically richer and better educated, and also in terms of age, with participants often younger or middle aged rather than from older age groups (Wilson 2014).

In many ways, these findings mirror more general literature on both the digital divide (Hargittai 2001; van Dijk and Hacker 2003) and other forms of participation such as civic volunteering and participation in politics. This literature has offered two central reasons as to why demographics should impact on participation levels. First, characteristics such as wealth and education may correlate with internet skills and ability (Hargittai et al. 2014; Hargittai and Shaw 2014): and without the necessary skills, people cannot participate in the given activity. They may even correlate with outright ability to access the internet (Servon 2008). Second, there may be social and network effects that mean people who are better off and more educated socialize with other highly participative people: hence they come to perceive participation as more normal (Brady et al. 1995). These dynamics could, of course, be self-reinforcing, with participation progressively being socialized as more (or less) normal in different demographic groups, gradually creating considerable structures of unequal

power which overlay the construction of crowdsourced knowledge.

The clear findings on demographic inequality of participants have led to a third stream of research on crowdsourced knowledge websites which concerns the extent to which the data produced on these sites is also demographically biased, considering the makeup of its user base. This refers to the possibility not just that data is incomplete, but that this incompleteness is related to both geography and demographics. In particular, the concern has been that the knowledge contained within such websites is biased towards the richer, better educated and more prosperous groups which also form their contributor base (a charge, we should add, which has also been levelled against traditional sources of information production—see e.g. Graham et al. 2014). This critique is arguably more fundamental than the possibility that crowdsourced knowledge is inaccurate, as it suggests that those basing their decisions on such knowledge might unwittingly be building in a demographic bias which in itself reinforces the structures of power which created these biases in the first place.

A wide variety of existing studies have documented that there do indeed appear to be demographic biases in crowdsourced data (Arsanjani et al. 2015; Girres and Touya 2010; Haklay 2010; Helbich et al. 2012), leading to what Graham et al. have described as "uneven geographies of user-generated information"(Graham et al. 2014). However, some research on the specific location of contributors has also highlighted how, in projects such as Wikipedia, non-locals can fill in information which has not been created by people within the specific area (Sen et al. 2015). This possibility of "non-local" knowledge contribution is intriguing; because it highlights that there must be a geographical scope within which demographic characteristics matter. That is to say, if crowdsourced knowledge concerning one poor area remains incomplete, perhaps a neighbouring area which is more privileged can fill it in. However, while we have some *prima facie* evidence to suggest this might be the case, no study has yet tested systematically whether these "neighbourhood effects" exist. The aim of this paper is to remedy this deficit.

Our empirical investigation is focused on OpenStreetMap [OSM], a crowdsourced knowledge site focused on the collection and sharing of volunteered geographic information (Elwood 2008; Elwood et al. 2013). OSM is a good choice for our purposes, for two main reasons. First, the data within it can also easily be attached to a geographic area, making it possible to quantify the extent to which different regions have complete data, and hence observe potential neighbourhood effects. Second, the data within it is rich yet, as previous studies have documented, shows variation in quality; and these variations have in the past been linked to demographic factors (Haklay 2010; De Sabbata et al. 2016). We will explore each of these factors in turn below, and hence highlight the particular types of demographic we will investigate in the empirical section.

First, the wealth of an area has been shown to correlate with its relative completeness on OSM (Mashhadi et al. 2015). Empirical studies have shown that a higher level of OSM data quality is more likely to appear in rich, advantaged urban areas than rural areas (e.g. Arsanjani et al. 2015; Girres and Touya 2010; Glasze and Perkins 2015; Haklay 2010; Helbich et al. 2012; Zielstra and Zipf 2010). For example, Haklay (2010) demonstrated that while the centres of big cities in England were well mapped, many parts of rural areas were not particularly well covered. Similarly, Girres and Touya (2010) reported that territories were well represented in rich areas in France while completeness became problematic in rural areas.

Second, education has also been found to play a role (though clearly it is also important to acknowledge that wealth and education are typically highly correlated). For instance, Budhathoki and Haythornthwaite (2013) reported that OSM contributors are highly educated, finding that 78% of the participants held a university degree. Stephens (2013) had similar findings. Similarly, several studies have also reported a positive relation between relevant skills and OSM contributions. Glasze and Perkins (2015) suggested that many of the mappers who made sustained contributions to the OSM project tended to have technical skills, while Yang et al. (2016) also reported that in their case study, most major OSM contributors in Germany, France, and the UK tended to have both rich experience in geographical data editing and a good level of technical skills.

Third, a variety of authors have reported a gender imbalance. For example, Haklay and Budhathoki (2010) conducted a survey for the OSM contributors, and reported that they are predominantly male: just 2.7% of the respondents were women. Similarly,

Schmidt et al. (2013) found that the majority of the active OSM contributors were male (96.2%). Lechner (2011) reported that amongst the OSM community participants, only about 2% were female, while Stephens (2013) also found out that male Internet users were significantly more likely to have heard of, used, and contributed to OSM than female Internet users. The reasons behind this gender imbalance have not yet been elaborated sufficiently (though see Steinmann et al. 2013).

Finally, two studies have documented the existence of digital divide between older and younger people. In a study of OSM contributors (Haklay and Budhathoki 2010), about two thirds of the contributors were found to be between 20 and 40 years old. 21.3% were in the range of 41–50 years, 10.4% were above 50 years old, and only 3.8% were below 20 years old. Girres and Touya, meanwhile (Girres and Touya 2010), have demonstrated that administrative regions were best represented in the areas with a young population in French OSM data.

In addition to demographic factors, it is worth highlighting that the size of population in an area has been considered a factor that might affect data quality on OSM, since the number of contributors in an area is presumed to be related to the number of objects digitally mapped in that area (Girres and Touya 2010). Indeed, empirical studies usually found a higher level of data quality in highly populated urban areas. For example, Helbich et al. (2012) stated that the OSM areas of highest accuracy were primarily highly populated urban areas, leading to the conclusion that these areas are subject to a higher validation rate as well as a higher correctness rate.

Also, several studies have reported a positive association between population density and OSM data quality. Mashhadi et al. (2015) reported that the higher the population density of an area, the higher the completeness of "point of interest" data. Dorn et al. (2015) also found that the more densely populated areas tended to have both higher coverage and higher thematic accuracy in OSM land use data. Nevertheless, results were not always consistent. For example, Ciepłuch et al. (2010) demonstrated that large sections of one of a selected group of towns were completely unmapped despite high population density. Dorn et al. (2015) also noted that although a high population density, as present in urban areas, tended to indicate a higher level of completeness, a low population density did not necessarily denote low completeness. Population density relates, of course, to a rural–urban divide, which has also been shown to have an impact. Studies have reported a higher level of completeness (e.g. Zielstra and Zipf 2010) as well as a higher level of topological accuracy (e.g. Helbich et al. 2012) in urban areas. Research has also reported significant differences in data accuracy and coverage within selected metropolitan areas (Arsanjani et al. 2015) and within a well mapped medium size city (Helbich et al. 2012).

Methodology

To restate, our research question is: do "neighbourhood effects" moderate geodemographic biases in crowdsourced knowledge websites? This is a question we assess in two parts. First, we offer our own measurement of the extent of geodemographic biases within OpenStreetMap [OSM]. Following the review of the literature above, we consider in particular biases related to wealth, education, gender and age, whilst also considering population size and density as control variables. Secondly (and more importantly), we assess whether any biases we can measure are ameliorated by "neighbourhood effects", by which we mean a dynamic whereby information in an area with few participants is filled in by neighbouring areas with more participants.

The empirical component of our paper is based on examining the extent to which a large, randomly sampled dataset of real world "points of interest" exists on OSM (in OSM terminology, a point of interest is a geographical feature on a map which exists in a single place: for example, a building). This dataset is drawn from a list of licenses for the sale of alcohol in the UK, made available by a consortium of local government actors.[1] This dataset is a useful choice for point of interest validation because, in theory at least, it ought to have a high level of accuracy: venues wishing to sell alcohol have a legal obligation to register their premises. Furthermore, while not a complete record of all different types of point of interest, alcohol licenses are nevertheless distributed to a wide variety of locations: not just bars, restaurants

[1] See: http://schemas.opendata.esd.org.uk/PremisesLicences.

and shops but also hospitals, schools, churches, businesses, hotels and even parks often have a license to sell alcohol. Hence the license registry contains a slice of validation data which is both varied and theoretically highly accurate.

We chose to take a random 5% sample of the license dataset, stratified by local authority, which equated to just over 2000 licenses. This gave us a sample large enough both to validate the overall accuracy of OSM and also to investigate geographic variation in this accuracy. The locations from which the licenses were taken are shown in Fig. 1. The amount of licenses per area varies, obviously, according to the density of licenses, with more being found in high density urban areas.

The dataset contained licenses which were valid at the time of data collection and also licenses which had subsequently been withdrawn, revoked or suspended (for example, if the location in question had closed down). We chose however not to exclude withdrawn licenses, as they allow us to also give an impression of the extent to which OSM responds to locations being shut down (arguably this is a difficult type of crowdsourced information to get right, as it involves not only people uploading points of interest to OSM, but checking existing ones against their local knowledge and potentially removing them if they are no longer valid). This is not a perfect check of course, as the fact that a license has been suspended does not mean the location itself has been shut down: a shop might decide to stop selling alcohol, for instance, but continue to sell other items. Nevertheless, it is the best data we have for addressing this question.

The sample of licenses was divided into two parts; each part was assigned to one of the authors in order to perform the validation. When checking whether a license existed on OSM, we simply navigated to the address of that license on the map, and checked to see if a corresponding point of interest was present with the same name. Some variation in names was allowed: for example, a restaurant called "John's Grill", in the same location as a license for "John Smith's Grill and Fish Bar", would be accepted as a match. There was also a slight tolerance in terms of location (indeed, as many points of interest in OSM do not have precise house numbers attached to them, it was not possible to assess the precise positional accuracy of the data). However, in general the aim was to be strict in terms of only accepting as "matches" points of interest which were clearly related to a license in the data. Once this process was complete the third author of the article was assigned 200 licenses at random, 100 each from the two tranches of data. They double coded these licenses, in order to provide a calculation of intercoder reliability. This triple coder

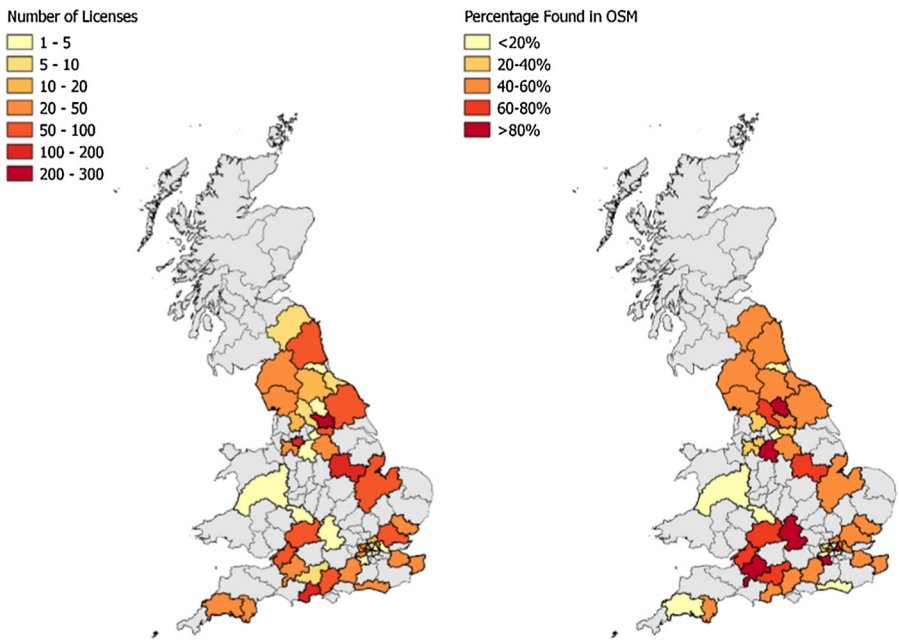

Fig. 1 Number of licenses in the sample (*left panel*) and % of found licenses (*right panel*), by postcode area. *Grey areas* indicate no data

agreed with the initial coding 85% of the time, which resulted in a Krippendorff's alpha of 0.69.[2]

Independent variables for the study were taken from the 2011 UK Census. The census was chosen because it offers data on the independent variables of interest which is both reliable, complete and quite granular. This allows us to assess the effect of demographics at different scales of aggregation. In particular, we chose to make use of three levels of aggregation (see Table 1): postcode "areas", which are large and typically contain hundreds of thousands of people; postcode "districts", which are medium sized areas and typically contain tens of thousands of people; and postcode "sectors", which are small areas which typically contain either hundreds or thousands of people. As the terminology of area, district and sector is not that helpful in terms of distinguishing their size, we will refer to these levels of aggregation as "large", "medium" and "small" throughout the rest of the paper.

The census provides information on all of our relevant independent variables of interest for all of these levels of aggregation. First, it provides the number of people in each area who belong to different "National Statistics Socio-economic Classes", which is an ordinal scale which groups people according to their profession, where professions broadly correlate to individuals' average earnings (see Rose and Pevalin 2003). This data source was combined into an average socio-economic class level for each area, which we take to represent the area's average wealth level. Second, it provides information on the number of people in each area who had achieved a certain level of education: again, this is an ordinal scale which was combined into an average education level for each area. The census also provides information on the number of males and females in an area, the average age of people living in an area, the number of people living there and the density of the population.

[2] Though the disagreement was not extensive, it is nevertheless worth reflecting on its potential causes. Several factors can be highlighted. First of all, the address data for licenses was not always particularly useful in finding the license on the map: some licenses did not come with a house number, just the name of a street, which at times could be very long. Second, as mentioned above, changes of names could produce the necessity for some judgment calls. Finally, the triple coding took place some months after the original coding had taken place: it is possible that some points of interest on OSM had been updated in this time.

Results

Initial descriptive results are presented in Table 2. In total 951 of the 2088 licenses in the ground truth dataset were found in OSM. When the stratified design is taken into account, our estimate for the completeness of this type of data in OSM data is just over 46%. As described above however, our sample contained both valid and invalid licenses. We found 49.1% of the valid licenses and 32.3% of the invalid licenses. This suggests that OSM is more likely to contain actually existing points of interest; but it may also include some points of interest which are no longer active (of course, we should highlight again that not all inactive licenses necessarily refer to premises which have closed).

It is worth considering the impact of the recency of the license on its probability of being found. Arguably, locations which have been around for longer ought to be more likely to be discovered. This is something that is examined in Fig. 2, which shows the percentage of licenses found by year of issue date, with confidence intervals. We can see that there is little evidence of a statistically significant difference between years with the exception of 2015 which is markedly lower (and which was the year of data collection). This seems to suggest that there is a time lag in terms of OSM being updated, but not a great one.

We will now move on to the analytical section of the paper. As we describe above, the main aim is to explore the extent to which neighbourhood effects exist in this point of interest data. However, before doing that, we need to offer a general measure of the extent to which demographic biases affect the completeness of our POI data.

We begin by investigating how the completeness of a given area correlates with its demographic characteristics in one local case study, Leeds. This was selected as a case study as it is the area with largest amount of data. Figure 3 shows the relationship between the relative level of wealth in a given area of Leeds and the amount of licenses found (it also shows the relationship between the wealth and education variables: as we might expect they are highly correlated). A positive correlation is visible between wealth and the percentage of licenses found however

Table 1 Types of postcode area

Level of aggregation	Postcode type	Average population	Minimum	Maximum
Large	Postcode area (e.g. OX)	721,400	18,320	1,359,000
Medium	Postcode district (e.g. OX1)	29,410	414	92,000
Small	Postcode sector (e.g. OX1 1)	7068	126	20,010

Table 2 Percentage of licenses found

	Total checked	Number found	Estimated completeness (%)	95% confidence interval lower bound (%)	95% confidence interval upper bound (%)
All licenses	2088	951	46.1	44.1	48.1
Valid licenses	1730	837	49.1	46.8	51.3
Invalid licenses	358	114	32.3	27.7	36.9

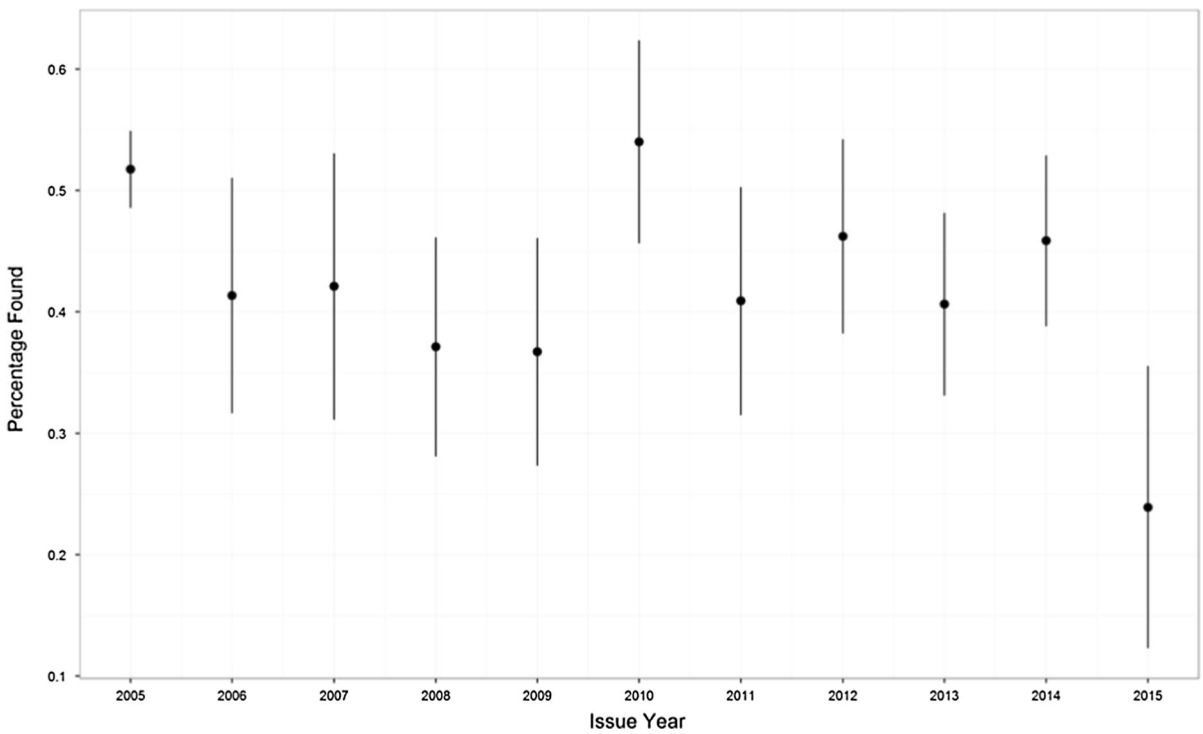

Fig. 2 Probability of a license being found related to issue date. *Bars* indicate 95% confidence intervals

also one with a high degree of residual error: a high level of completeness can be found in areas classified by the census as of low wealth, and vice versa low level of completeness can be found in areas classified as relatively high wealth. Similarly, values of completeness are not spatially auto-correlated, and can change dramatically from one postcode sector to the next, as shown in Fig. 4. This could be a sign of zoning effects, or could combine with the effect of specific user behaviour in the area.

We will now move on to address more systematically the question of geodemographic biases in the

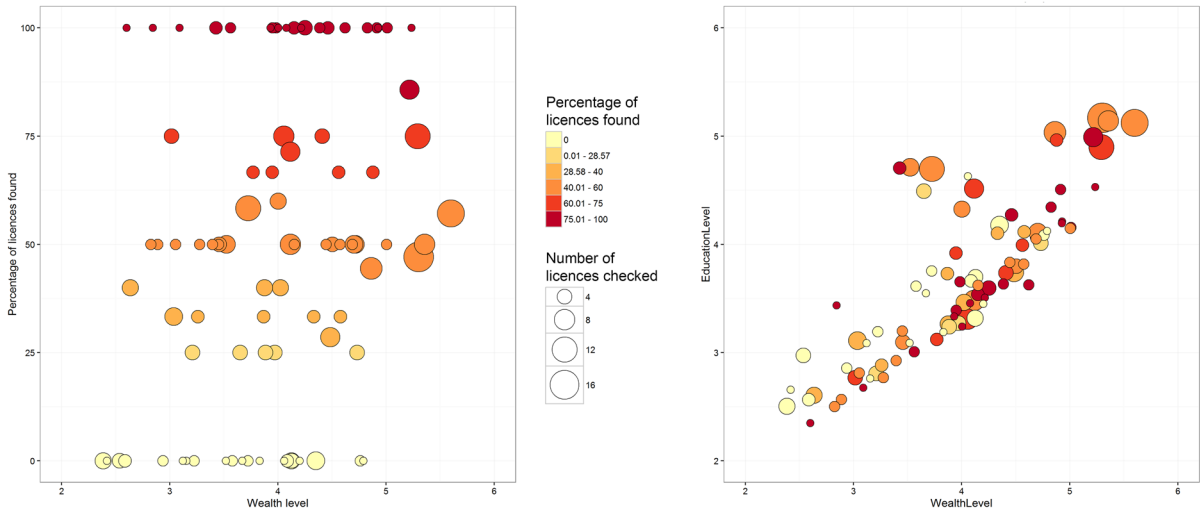

Fig. 3 Relationship between wealth of an area and percentage of licenses found (*left panel*), and relationship between wealth and education of an area (*right panel*)

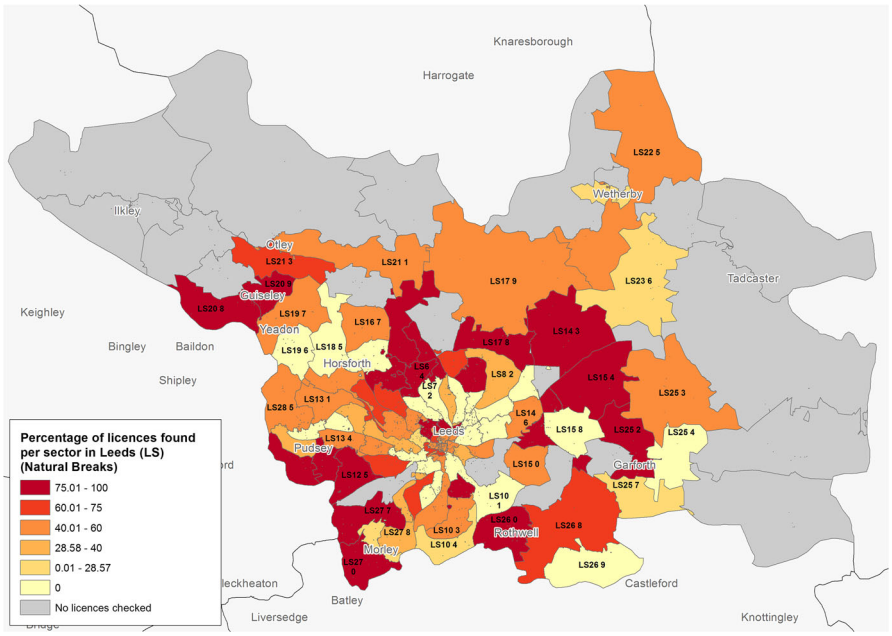

Fig. 4 Percentage of licenses found in different areas of Leeds

data. We tackle this question with a series of logistic regressions, shown in Table 3. Each regression seeks to assess the factors correlated with the likelihood of an individual license being found in OSM, given the characteristics of the area within which it finds itself. While our interest is, overall, in completeness at the aggregate level, we chose to model effects at the micro level rather than aggregate to the area level because our data is at this micro level.

We estimate three groups of two models; each of the three groups looks at one potential level of aggregation of our independent variables of interest. Hence, for example, in the first line of Model 1.1, the effect of interest is the education level of the large

Table 3 Logistic regression models explaining the likelihood of a license being found in OSM

	Large area		Medium area		Small area	
	Model 1.1	Model 1.2	Model 1.3	Model 1.4	Model 1.5	Model 1.6
Intercept	0.01***	0.00***	0.09***	0.04***	0.12***	0.03***
Wealth level	2.49***		1.78***		1.59***	
Education level		3.33***		2.08***		1.89***
Number of males	1.09**	1.09**	1.14	0.97	0.86	0.64
Mean age	1.03	1.04	0.99	1.00	0.99	1.01
Population density	0.98***	0.97***	0.99**	0.99***	0.99*	0.99***
Population size	0.96**	0.96**	0.94	1.02	1.08	1.27
Valid license?	2.83***	2.89***	2.39***	2.42***	2.35***	2.43***
Creation year	0.95***	0.95***	0.95***	0.95***	0.96***	0.96***
Observations	1875	1875	1859	1859	1825	1825
AIC	2522	2525	2486	2495	2438	2450

Coefficients are exponentiated

Population size and number of males measured in 1000 s

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

postcode area within which the license finds itself. In the first line of model 1.5, by contrast, the education level of the small postcode sector is what is considered. We have two sub-models within each of our levels of aggregation because we wanted to assess the impact of wealth and education level in separate models, as these two indices are highly correlated as we saw above. The coefficients reported are exponentiated, which means that they can be interpreted as a percentage effect on the underlying chance of the license being discovered. Hence, for example, in model 1.1, we can see that for every additional 1000 males in an area, the chance of a license being discovered increases by 9%.

The models in the table support some of our initial hypotheses about demographic biases in the data. Wealth level correlates strongly with the probability of a license being discovered: increasing this level by a point increases the probability of discovery by around 150% when considered at the largest level of aggregation (model 1.1), with this effect declining but remaining strong and positive for the medium and small levels. The story is similar for the education variable: a strong positive effect at the large area of aggregation which declines, but remains strong and positive, and the medium and small levels. By contrast, we find only limited support for our gender hypothesis: increasing numbers of males in an area increases the probability of a license being found if we look at the largest level of aggregation, but has no effect at any other level. There are, meanwhile, no results for our age variable. In terms of our control variables, at the large level, increases in population made a license less likely to be found; increases in population density, meanwhile, made it less likely that licenses would be found at all levels. Meanwhile, our control variables for the validity of the license is strong and consistently significant across all models, as we would expect.

The comparison between the models, with stronger effects typically found at the large area level, seems to suggest that the broad geographical location a license is located in matters more than its immediate vicinity. This would suggest, in other words, that users contribute content from their wider region, rather than just their immediate vicinity, and would lend support to the idea of a neighbourhood effect. However direct comparison between the models is also difficult, because the amount of observations and the variance of independent variables changes.

Hence we also produce a second set of models, found in Table 4. These models directly address the question of whether neighbourhood effects exist. We produce two sets of models, one looking at the medium level of aggregation and one looking at the small area of aggregation. As well as looking at the impact of the variables already mentioned, these models also include the average wealth and educational level of neighbouring medium and small areas (we do not look at the average age or the number of

Table 4 Logistic regression models explaining the likelihood of a license being found in OSM, taking into account neighbourhood effects

	Medium area		Small area	
	Model 2.1	Model 2.2	Model 2.3	Model 2.4
Intercept	0.00***	0.01**	0.00*	0.00**
Wealth level	8.32***		15.23*	
Education level		2.21[a]		18.98**
Number of males	1.00*	0.99***	0.99**	0.99***
Mean age	1.16	1.28	0.93	1.04
Population density	0.74	0.62	1.15	0.93
Population size	0.98*	1.01	0.99	1.01
Valid license?	2.56***	2.57***	2.45***	2.40***
License creation year	0.95***	0.95***	0.95***	0.96**
Neighbouring wealth level	8.25***		11.63[a]	
Neighbouring education level		1.75		8.98 [a]
Interaction term	0.63***	0.91	0.57[a]	0.54*
Observations	1859	1859	1748	1748
AIC	2485	2485	2329	2336

Coefficients are exponentiated

Population size and number of males measured in 1000 s

[a] $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

males in neighbouring areas, as these variables did not produce strong evidence in our first set of models).[3] These averages are weighted to take into account the population of these neighbouring areas (so having a more populous wealthy neighbouring area counts for more than a less populous wealthy one). If neighbourhood effects exist, we would expect the chance of discovering a particular license to increase as the wealth or education of neighbouring areas goes up. Furthermore, the term for the neighbouring areas and the job and education level of the area itself are interacted, because we would expect neighbourhood effects to vary with the wealth or education of the given area: if the area itself is wealthy, the characteristics of its neighbours matter less.

The results provide support for the idea that neighbourhood effects do exist, though the results for models looking at wealth are stronger than those looking at education. The coefficients of the new variables of interest in these models point in the expected direction: increasing wealth and education in neighbouring areas does have a positive effect on the chances of a license being found. However, the effect is only statistically significant in model 2.1 (though it is on the borderlines of significance in model 2.3 and model 2.4). The coefficient of the interaction term also goes in the expected direction, reducing the chance of a license being found. This has the effect of moderating the impact of increases in neighbouring area wealth or education: as an area gets wealthier or better educated, the impact of the characteristics of its neighbours diminishes.

The impact of the interaction term in the model is easier to interpret in a visualisation, presented in Fig. 5. This figure shows post-estimations from model 2.1, showing how the probability of observing a license changes as the wealth of neighbouring areas increases. The different lines represent different levels of wealth level in the actual area within which the license finds itself: the red line is fixed at the first quartile, while the blue line is fixed at the third quartile. For areas which are at the first quartile for wealth level, there is a positive slope: i.e. increasing wealth in neighbouring areas increases the chance of the license being discovered. For areas with a level of wealth in the third quartile, by contrast, the line is flat: increases in the wealth of neighbouring areas make no difference. Furthermore, the figure also reveals that the size of the neighbourhood effect, while modest in absolute terms, is relatively speaking quite considerable. The confidence intervals overlap at the right hand edge of the graph, which indicates that a poor area

[3] Neighbouring status is identified through use of the postcode structure, hence, for example, the area of aggregation OX1 1 is considered to "neighbour" all other areas within OX1.

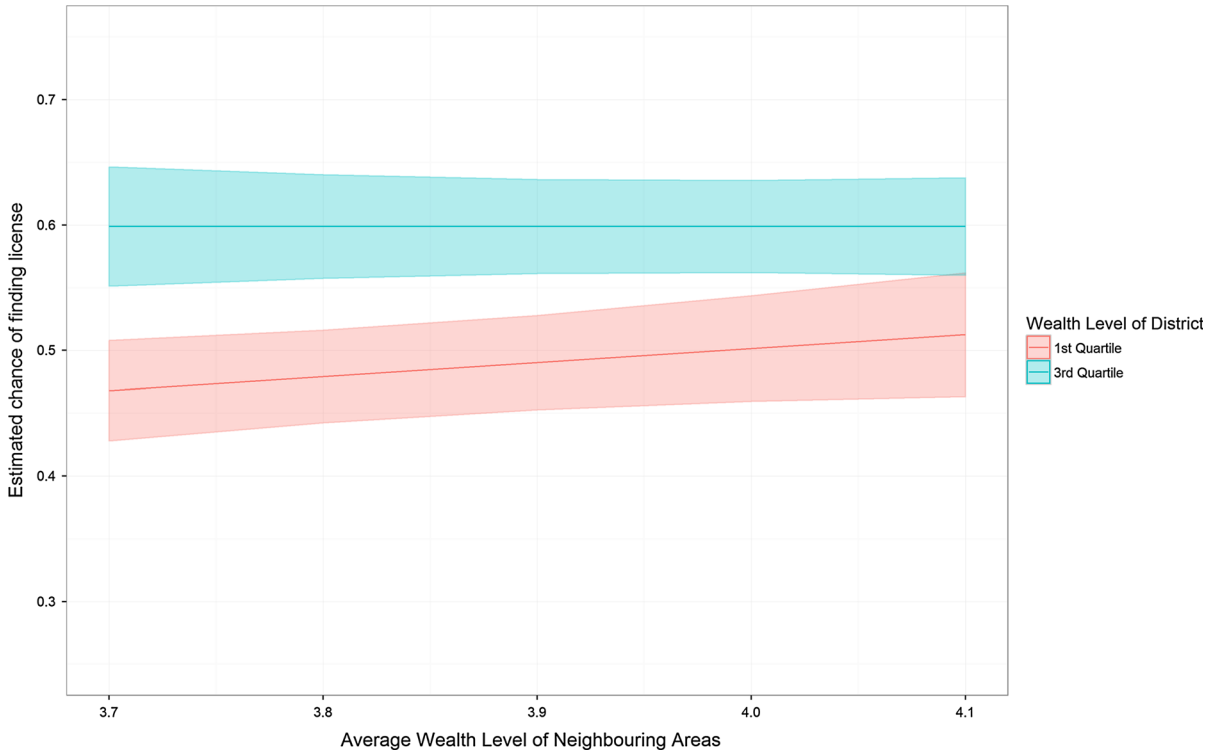

Fig. 5 Post-estimations from model 2.1. Average neighbouring wealth level varies between its first and third observed quartiles. All other variables held at their mean value

surrounded by rich areas is statistically indistinguishable from a rich area in terms of chance of license data being present.

Discussion and conclusions

This paper has sought to further investigation into geodemographic biases in crowdsourced knowledge websites. It has made two main contributions. First, making use of a novel dataset of point of interest data, it has shown that OpenStreetMap is around 50% complete, and that this completeness is related to the demographic characteristics of the area to which the dataset relates. Second, and more importantly, it has shown systematic evidence of neighbourhood effects, whereby lack of contributions in one area can be compensated by contributions from another area. This neighbourhood effect was relatively strong, to the extent that poor areas surrounded by rich areas were essentially the same as rich areas themselves.

What do these results indicate for the future of crowdsourced knowledge websites? On the one hand, they provide some reasons to be optimistic: while geodemographic inequality in contribution levels might be relatively intractable, the websites themselves may continue to get more complete as those people who do contribute fill in the blanks for those that do not. Hence, geodemographic inequality need not be a fatal objection to crowdsourced knowledge. However, there are also important reservations to this position. First, it is unclear whether the replacement data being contributed by other participants is truly of the same quality as data which could have been contributed by those who actually live in an area: the type of data being brought in may also reflect participants' own biases and interests. Second, neighbourhood effects could end up masking geodemographic bias in contributor bases, which might end up giving people the perception that websites are more inclusive than they actually are. For both these reasons, we should be cautious about celebrating the existence of neighbourhood effects, even while

acknowledging their potential impact. Further research on the exact types of information contributed by neighbouring participants would be useful to further our understanding of their impact.

We should conclude by highlighting the limitations in the paper, and thus point the way for further research. Firstly, we made use of a limited point of interest dataset of alcohol licenses. Such points of interest cannot be assumed to represent the whole of OSM, and if we had looked at other types of map feature we might have seen different results. Second, we make use of area level proxy measures for our demographic variables: however in a certain sense this risks ecological fallacy, as of course a deprived area may well still contain enough rich and well educated individuals to produce sufficient levels of contributors. Finally, of course, we look only at the UK context: a study of another country might again have shown different results. Further research on these aspects would be welcome and expand our knowledge of the extent of geodemographic biases in crowdsourced knowledge websites.

Compliance with ethical standards

Conflict of interest We have no conflicts of interest.

Human and animals rights The research did not involve human participants or animals.

Informed consent The research did not involve the collection of any personally identifying data, nor anything that could be construed as identifying an individual, hence informed consent was not relevant.

References

Arsanjani, J., Mooney, P., Zipf, A., & Schauss, A. (2015). Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In J. Arsanjani, A. Zipf, P. Mooney & M. Helbich (Eds.), *OpenStreetMap in GIScience*. Zurich: Springer.

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., et al. (2014). Citizen science: Next steps for citizen science. *Science, 343*(6178), 1436–1437.

Brady, H. E., Verba, S., & Schlozman, K. L. (1995). Beyond SES: A resource model of political participation. *American Political Science Review, 89*(2), 271–294.

Budhathoki, N. R., & Haythornthwaite, C. (2013). Motivation for open collaboration: Crowd and community models and the case of OpenStreetMap. *American Behavioral Scientist, 57*(5), 548–575.

Ciepłuch, B., Jacob, R., Mooney, P., & Winstanley, A. C. (2010). Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In *Proceedings of the 9th international symposium on spatial accuracy assessment in natural resuorces and enviromental sciences 20–23 July 2010*. University of Leicester.

Cobo, C., Bulger, M. E., Bright, J., & den Rooijen, R. (2016). What role do "power learners" play in online learning communities? In *Proceedings of LINC, 7th conference of the learning international networks consortium* (pp. 83–92).

De Sabbata, S., Tate, N., & Jarvis, C. (2016). Characterizing volunteered geographic information using fuzzy clustering. In *Proceedings of the 9th international conference on geographic information science*. Canada: Montreal.

Dorn, H., Törnros, T., & Zipf, A. (2015). Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany. *ISPRS International Journal of Geo-Information, 4*(3), 1657–1671.

Elwood, S. (2008). Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal, 72*(3–4), 173–183.

Elwood, S., Goodchild, M. F., & Sui, D. (2013). Prospects for VGI research and the emerging fourth paradigm. In Crowdsourcing Geographic (Ed.), *Knowledge* (pp. 361–375). Dordrecht: Springer.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy, 43*(1), 1–20.

Girres, J., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS, 14*(4), 435–459.

Glasze, G., & Perkins, C. (2015). Social and political dimensions of the OpenStreetMap project: Towards a critical geographical research agenda. In J. Arsanjani, A. Zipf, P. Mooney & M. Helbich (Eds.), *OpenStreetMap in GIScience*. Zurich: Springer.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Goode, L. (2009). Social news, citizen journalism and democracy. *New Media and Society, 11*(8), 1287–1305.

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers, 104*(4), 746–764.

Haklay, M. (2010). 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.

Haklay, M. (2016). Why is participation inequality important? In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann & R. Purves (Eds.), *European handbook of crowdsourced geographic information*. London: Ubiquity Press.

Haklay, M., & Budhathoki, N. (2010). OpenStreetMap–Overview and motivational factors. In *Proceedings of the horizon infrastructure challenge theme day, University of Nottingham*.

Hargittai, E. (2001). Second-level digital divide: Mapping differences in people's online skills. *First Monday, 7*(4).

Hargittai, E., Connell, S., Klawitter, E. F., & Litt, E. (2014). Persisting effects of internet skills on online participation. *2014 TPRC conference paper*. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2418033. Accessed 22 May 17.

Hargittai, E., & Shaw, A. (2014). Mind the skills gap: The role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, Communication and Society, 18*(4):424–442.

Helbich, M., Amelunxen, C., & Neis, P. (2012). Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata. In *Proceedings of GI_forum* (pp. 24–33).

Hill, B. M., Shaw, A., Cohen, N., Chang, L., Krosnick, J., Valliant, R., et al. (2013). The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS ONE, 8*(6), e65782. doi:10.1371/journal.pone.0065782.

Jönsson, A. M., & Örnebring, H. (2011). User-generated content and the news: Empowerment of citizens or interactive illusion? *Journalism Practice, 5*(2), 127–144.

Lechner, M. (2011). *Nutzungspotentiale crowdsource-erhobener Geodaten auf verschiedenen Skalen*. Freiburg: Albert-Ludwigs University Press.

Mashhadi, A., Quattrone, G., & Capra, L. (2015). The impact of society on volunteered geographic information: The case of OpenStreetMap. In J. Arsanjani, A. Zipf, P. Mooney & M. Helbich (Eds.), *OpenStreetMap in GIScience*. Zurich: Springer.

Panciera, K., Halfaker, A., & Terveen, L. (2009). Wikipedians are born, not made: A study of power editors on wikipedia. In *Proceedings of the ACM 2009 international conference on supporting group work* (pp. 51–60).

Rafaeli, S., & Ariel, Y. (2008). Online motivational factors: Incentives for participation and contribution in wikipedia. In *Psychological aspects of Cyberspace: theory, research, applications* (pp. 243–267). Cambridge, UK: Cambridge University Press.

Rose, D., & Pevalin, D. J. (2003). *A researcher's guide to the national statistics socio-economic classification*. Thousand Oaks: Sage.

Sauermann, H., & Franzoni, C. (2015). Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences, 112*(3), 679–684.

Schmidt, M., Klettner, S., & Steinmann, R. (2013). Barriers for contributing to VGI projects. In *Proceedings of ICC* (Vol. 13).

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., et al. (2016). Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information, 5*(5), 55.

Sen, S. W., Ford, H., Musicant, D. R., Graham, M., Keyes, O. S. B., & Hecht, B. (2015). Barriers to the localness of volunteered geographic information. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems—CHI'15*, New York (pp. 197–206). New York, USA: ACM Press.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science, 31*(1), 139–167.

Servon, L. J. (2008). *Bridging the digital divide: Technology, community and public policy*. Malden: Blackwell.

Shirky, C. (2008). *Here comes everybody. The power of organization without organizations*. London: Allen Lane.

Steinmann, R., Häusler, E., Klettner, S., Schmidt, M., & Lin, Y. (2013). Gender dimensions in UGC and VGI: A desk-based study. In *Proceedings of GI_Form 2013 creating the GISociety* (pp. 355–364).

Stephens, M. (2013). Gender and the GeoWeb: Divisions in the production of user-generated cartographic information. *GeoJournal, 78*(6), 981–996.

Van Dijk, J., & Hacker, K. (2003). The digital divide as a complex and dynamic phenomenon. *The Information Society, 19*(4), 315–326.

Voigt, C., & Bright, J. (2016). The lightweight smart city and biases in repurposed big data. In *The 2nd international conference on human and social analytics (HUSO 16)*.

Wilson, J. (2014). Proceed with extreme caution: Citation to wikipedia in light of contributor demographics and content policies. *Vanderbilt Journal of Entertainment and Technology Law, 16*(4), 857–908

Yang, A., Fan, H., & Jing, N. (2016). Amateur or professional: Assessing the expertise of major contributors in OpenStreetMap based on contributing behaviors. *ISPRS International Journal of Geo-Information, 52*(2), 21.

Zielstra, D., & Zipf, A. (2010). A comparative study of proprietary geodata and volunteered geographic information for Germany. In *13th AGILE international conference on*. Guimarães.