Health & Place 50 (2018) 130–136

Contents lists available at ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

OpenStreetMap data for alcohol research: Reliability assessment and quality indicators

Jonathan Bright[a,*], Stefano De Sabbata[a,b], Sumin Lee[a], Bharath Ganesh[a], David K. Humphreys[c,d]

[a] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, United Kingdom
[b] School of Geography, Geology and the Environment, University of Leicesteró, Leicester LE1 7RH, United Kingdom
[c] Department of Social Policy and Intervention, University of Oxford Barnett House, 32 Wellington Square, Oxford OX1 2ER, United Kingdom
[d] Green Templeton College, University of Oxford, 43 Woodstock Road, Oxford OX2 6HG, United Kingdom

ARTICLE INFO

Keywords:
Alcohol availability
Alcohol licensing
Alcohol-related harm
OpenStreetMap
Big data

ABSTRACT

There is a growing interest in using OpenStreetMap [OSM] data in health research. We evaluate the usefulness of OSM data for researching the spatial availability of alcohol, a field which has been hampered by data access difficulties. We find OSM data is about 50% complete, which appears adequate for replicating findings from other studies using alcohol licensing data. Further, we show how OSM quality metrics can be used to select areas with more complete alcohol data. The ease of access and use may create opportunities for analysts and researchers seeking to understand broad patterns of alcohol availability.

1. Introduction

Controlling the physical availability of alcohol is widely regarded to be a key strategy for helping national and local policymakers prevent alcohol misuse and its various related harms (Babor et al., 2010). The link between physical availability of alcohol and harm is supported by a growing body of international research showing a statistical association between the spatial availability of alcohol (i.e. alcohol outlet density) and the prevalence of a range of acute and chronic physical and social harms (Bryden et al., 2012; Campbell et al., 2009; Holmes et al., 2014; Popova et al., 2009). To date, this literature has been dominated by studies from only a small number of geographic contexts, such as the United States, Canada, Australia and Scandinavia. Until recently, studies from the UK have been underrepresented in this literature (Holmes et al., 2014; Humphreys and Smith, 2013).

To assess the spatial availability of alcohol, it is necessary to collect data on the geographic location of retail alcohol outlets. In the UK, as in many other countries, relevant data concerning the location and trading practices of alcohol outlets are usually found in local government licensing records (Hadfield, 2006; Humphreys and Eisner, 2010). In England and Wales, the conditions of the Licensing Act (2003) stipulate that these records should be made openly available to members of the public (HM Government, 2012; HMSO, 2005). As a consequence, gaining access to these administrative records for the purpose of generating small area measures of alcohol availability should be relatively straightforward. But unfortunately this is not the case, as was found in Humphreys and Smith (2013) who investigated the barriers to accessing and using licensing records across Greater London (n = 33 administrative areas). This study found that less than a third of local authorities made licensing registers openly available to the public (i.e. via online registers)—requiring specific data requests to be made to the local authority. While most (60%) licensing authorities were willing to provide the requested records under the Freedom of Information Act, the data received differed considerably between areas in terms of: data format (e.g. file types); content (e.g. variables); accuracy and completeness. The inconsistency of these data made the costs of data cleaning and reformatting prohibitively expensive (Humphreys and Smith, 2013).

Since this study was published there have been several notable studies examining the relationship between the spatial availability of alcohol and a variety of physical and social harms in a UK context. In Scotland, Richardson et al. (2015) used data from Liquor Licensing Boards across 4 large cities (Glasgow, Edinburgh, Aberdeen and Dundee) to create measures of alcohol outlet density for both on- and off-premises alcohol consumption. In Wales, the CHALICE study used licensing data from 22 local authorities across Wales to calculate proximity measures of alcohol outlet density for residences across Wales over a six-year period (2005–2011) (Fone et al., 2016). In both cases, authors underlined the challenges of using administrative licensing records to generate spatial measures of alcohol availability. In the case of the Scottish study, Richardson et al. (2015) noted that inconsistencies in the arrangement of data between licensing liquor

* Corresponding author.
E-mail address: jonathan.bright@oii.ox.ac.uk (J. Bright).

https://doi.org/10.1016/j.healthplace.2018.01.009
Received 10 July 2017; Received in revised form 20 November 2017; Accepted 26 January 2018
Available online 12 February 2018

boards limited the scope of the study to four large cities due to resources that would be required to collect and prepare the data for analysis. Similarly, researchers from the CHALICE study documented the considerable effort and resources required to attain data from local authorities. On average 4 separate requests (min = 1, max = 13) to each local authority, often over 12 months or more were required to collect the relevant data. Consistent with the findings of Humphreys and Smith (2013), these studies found that the data eventually received varied considerably in format and accuracy (Fone et al., 2016; Fry et al., 2016).

In response to data access problems such as these, some researchers have taken an alternative approach to acquiring data to measure the spatial availability of alcohol. In a recent study documenting patterns of spatial availability of alcohol in England between 2003 and 2013, researchers from the University of Sheffield purchased market research data from commercial organisations specialising in the collection of data on the leisure industry (CGA Strategy and Nielsen) (Angus et al., 2017). This is the most comprehensive study to date to examine patterns of spatial alcohol availability in a UK context. In this study, researchers were able to calculate measures of availability for all English postcodes in terms of proximity (i.e. distances to the nearest outlet in metres) and density (i.e. no. of outlets within 1 km). The richness of the data allows researchers to differentiate between on- and off-premises alcohol availability, as well as between different categories of licensed premises (e.g. pubs, bars, nightclubs, restaurants, supermarkets, convenience stores, etc). While the use of market research data presents many new opportunities for generating rich measures of spatial alcohol availability, it is questionable whether it presents a sustainable solution to the problem of accessing and using data to generate spatial measures of alcohol availability. For example, purchasing these data can be extremely costly, making them too expensive for public health analysts and most other researchers to obtain.

In summary, there are significant barriers preventing researchers and practitioners from easily monitoring the nature and extent of the spatial availability of alcohol as well as evaluating the impact of changes to availability that may occur as a result of various policy interventions (Humphreys and Eisner, 2014; Humphreys and Smith, 2013). In this paper we seek to explore a potential solution to this problem, which is the possibility that data generated by the "volunteer geographic information" site OpenStreetMap can be used to generate measures of the spatial availability of alcohol which would be beneficial for researchers and practitioners interested in monitoring alcohol availability. Volunteer geographic information (VGI) refers to spatial information provided voluntarily by individuals to websites which are engaged in some sort of collaborative mapping project (Goodchild, 2007). VGI is becoming increasingly popular in health research in general, and as such a developing body of literature already exists pointing to its potential usefulness in areas where little alternative data is available (see Cinnamon and Schuurman, 2013; Stensgaard et al., 2009). Much of this work has looked at how the principles of VGI can be applied to research in public health through the use of bespoke applications (see Boulos et al., 2011 for a review), where patients, doctors or other concerned individuals are engaged in a collaborative act of data generation on a particular topic of interest such as public health surveillance. In these cases, VGI allows researchers to engage with communities and potentially produce robust datasets to highlight spatial relationships that may not have been previously evident (Goranson et al., 2012; Quinn and Yapa, 2015; Meenar, 2017; Pfeiffer and Stevens, 2015).

Alongside VGI initiatives specifically designed for health research, authors are also starting to point to how VGI that has been created in other projects might be repurposed to be used directly in health research (Mooney and Corcoran, 2011; Bergquist and Rinaldi, 2010). OpenStreetMap (OSM) is the most obvious example in this regard (cf. Goodchild and Li, 2012, 111). OSM is a collaborative mapping tool which seeks to create a complete map of the world through the contributions of its volunteers. It contains data on a variety of features which might be of interest to spatial health research: from the location of public amenities such as hospitals, schools and leisure centres to venues for the sale of food and alcohol.

In this paper, we address specifically the potential usefulness of OSM data for research on the spatial availability of alcohol. As a data source, OSM has numerous theoretical advantages over the available data sources for alcohol research described above: it makes its data freely available for download; the data comes in machine readable formats which facilitate its manipulation and use; it is continually updated by volunteers; and it contains granular data on different types of venue, allowing researchers to distinguish, for example, between off- and on-license sale. However, before OSM can be integrated into alcohol research, important questions about its completeness and accuracy need to be answered. There is an extensive existing literature which has assessed various elements of the accuracy of OSM (e.g., Arsanjani et al., 2015; Haklay, 2010; Helbich et al., 2012; Girres and Touya, 2010; Mashhadi et al., 2015; Senaratne et al., 2017; Zielstra and Zipf, 2010; Bright et al., 2017), which has been cautiously positive whilst also noting that the resource is not complete. However, the majority of these studies have addressed the completeness of the road network: to date no study has looked specifically at its potential to measure spatial availability of alcohol. Hence the question of the extent to which OSM can be useful for alcohol researchers remains unanswered.

The aim of this paper is to fill this gap, by directly addressing the question of whether OpenStreetMap is a useful proxy measure for the spatial availability of alcohol. We make two contributions in particular. First, we systematically validate the completeness of OpenStreetMap data for a variety of different types of alcohol point of sale, using both a large scale manual validation task and also comparing data from OSM with freely released data from a recent study of the spatial availability of alcohol in Scotland. Second, we describe how indicators of the quality of OSM data in a particular area can be constructed, and show how these can be used to select a subset of areas with more complete alcohol data. Accompanying the study, we also release a dataset containing indicators of alcohol outlet density in OpenStreetMap for all postcode sectors in England, Scotland and Wales, together with accompanying quality metrics.[1]

1.1. Study design

The study has two parts. (1) First, we perform a validation exercise to assess the accuracy of the alcohol point of sale data which can be extracted from OpenStreetMap. We make use of two "ground truth" datasets: a partial dataset of local alcohol licenses accessed from the Local Government Association (LGA) open data platform, and data released by a recent study of alcohol related harms in Scotland (Richardson et al., 2015). We measure the extent to which data in OSM duplicates the indicators found in both these datasets. (2) Second, we describe a mechanism for selecting areas of "high quality" OSM data, and explore the extent to which this mechanism can be used to select areas with more complete alcohol data.

1.2. Data

We make use of a variety of sources of data for the two steps of the study design described above. First, we collected a sample of premise licenses accessed through a LGA open data partnership as our ground truth dataset.[2] The LGA have recently introduced a number of resources to help promote open and transparent local government.

[1] This dataset can be found at http://researchdata.ox.ac.uk/.

[2] This dataset is available from: http://schemas.opendata.esd.org.uk/PremisesLicences.

Through this initiative local authorities are financially incentivised to upload licensing data on premise licenses for a small fee that is only redeemable if data are uploaded in a preferred format, consistent with a pre-specified schema provided by the LGA (and available from GitHub). At present licensing records available on this platform cover only a subset of areas in the UK.[3] We took a random 5% sample of this dataset, stratified at the local authority level (2088 license records in total). We should note 358 of these records were for "invalid" licenses, which had been issued then subsequently either expired or been revoked. Including these licenses also allows us to measure the extent to which OSM contains invalid alcohol data within it. In total, these 2088 license records formed the ground truth data for our validation study.

Second, we made use of data released by researchers at the Centre for Research on Environment, Society and Health (CRESH) at the University of Edinburgh (Richardson et al., 2015), which will hereafter be referred to as the CRESH study.[4] The data these researchers collected was obtained by contacting local licensing boards in Scotland in 2012. While the analysis in the CRESH study was focused on only four cities, the data release contains information on alcohol outlet density across all Scottish datazones (a small area spatial unit of the Scottish census). In particular, for each datazone the study authors measured the amount of qualifying outlets within an 800 m radius from the population-weighted centroid.[5] They also released data on the Standardised Mortality Ratio for each datazone (that is, the number of alcohol related deaths observed in each datazone divided by the average number which would be expected given the population of that datazone), and we supplemented this data with publicly available information from the 2001 Scottish census and the Scottish Index of Multiple Deprivation.

Third, data was collected from OpenStreetMap itself. To perform the validation task using the dataset of licenses, we simply made use of OSM's publicly facing web portal, which allows anyone to browse the map data contained within OSM.[6] To create our own measures of alcohol outlet density for comparison with the CRESH study we downloaded the complete OSM database from the website GeoFabrik[7] on June 4th, 2015. We then queried the database with a list of establishment types which we believed were likely to sell alcohol for both on and off site consumption. A full list of the search terms used in this query can be found in the appendix.

Finally, we also created an indicator of the quality of OSM data in a given area, inspired by the work of Barron et al. (2014) on quality measures. The indicator is an average of seven separate characteristics of a given area which can be calculated through data which OSM makes available for download. These characteristics are: the number of map "features" per square kilometre (a feature is anything that has been added to the map such as roads, houses, parks, etc.); the average number of "edits" per map feature (where edits reflect any change made to the map by a user); the number of users who made at least one edit in the area; the number of features edited by more than one user; the number of days with at least one edit in the area; the number of days since the last edit was made; and, the amount of buildings in the area which contain complete address information. The index was calculated on the full historical dataset for England and Scotland, downloaded from GeoFabrik. A more detailed description of how the quality indicator was created can also be found in the appendix.

[3] At the time of sampling the dataset covered 590 postcode districts from 48 local authority areas. These licenses were all issued in the period 2005 – 2015 inclusive.

[4] This dataset is available from: https://cresh.org.uk/webmap/about-creshs-map-of-neighbourhood-alcohol-and-tobacco-environments-in-scotland.

[5] A population-weighted centroid is the centre of a spatial unit that accounts for the spatial distribution of population within that unit. Population-weighted centroids were provided to the study by the National Records of Scotland.

[6] This portal can be found here: https://www.openstreetmap.org.

[7] Downloaded from: http://download.geofabrik.de/europe/great-britain.html.

1.3. Statistical analysis

We perform three main analyses. First, for each venue license in our LGA dataset, one of the authors of the paper navigated to the license address on the OpenStreetMap web portal, and recorded whether the equivalent venue also existed on OpenStreetMap. A further author of the paper was assigned 200 of the 2000 licenses at random, and repeated the procedure, which allowed us to assess inter-coder reliability of our validation exercise. This double coding resulted in an 85% agreement, which gave a Krippendorff's alpha of 0.69. The data generated by this process allowed us to estimate the overall percentage completeness of OpenStreetMap with respect to alcohol license data.

Second, we compared the measures of alcohol outlet density released by the CRESH study with the ones we generate using OSM data. We observed the correlation between the two measures using Pearson's R, and also attempted to replicate the regression analyses reported in the CRESH study. Our models, we should note, are not exact duplications: the dependent variable used in the CRESH study is a count of alcohol related harms, hence they employ a Poisson model. However they did not openly release these data due to privacy concerns (instead, they provide a Standardised Mortality Ratio band for each Scottish datazone as described above). Our model is hence an ordered logistic regression.

Finally, we assessed the extent to which the quality indicators described above can be used to find higher quality data on alcohol outlet availability. We do this by observing the extent to which the completeness of our OSM data, and the correlation between the CRESH data and our data, varies over different deciles of the quality indicator.

2. Results

2.1. Validating OSM Alcohol Sale Data

We begin by reporting the results of our validation exercise (see Fig. 1). In total, of the 1730 valid premises licenses we checked on OSM, 837 were found to be present. Once stratification is taken into account, this allows us to estimate the completeness of alcohol point of sale data in OSM at 49.1%. We can also assign a standard binomial proportion 95% confidence interval to this estimate, which runs from 46.8% to 51.3%. This indicates that OSM contains a significant amount of alcohol license data within it, though it is also far from complete.

Of the 358 invalid licenses, 32.3% of them were found to be present on OSM (95% confidence interval bounds run from 27.7% to 36.9%). This indicates that OSM may be slow to remove points of interest which have closed down and hence, in addition to containing only a partial record of actually existing alcohol points of sale, OSM may indicate the presence of some points of sale which no longer exist. Of course, it is worth noting that OSM does not seek to keep records of licenses but rather map features, broadly conceived, which may mean that from the point of view of the map, outright removal of a feature would not make sense. For example, a pub which has shut down still remains relevant to include on a map if the building which houses it has not actually been demolished.

It is worth considering how the accuracy of OSM breaks down by type of license. Hence we also coded each license in our dataset into one of five broad categories: bars and pubs, hotels, shops (which includes supermarkets), restaurants and a "miscellaneous" category (which includes licenses for churches, schools, hospitals and other points of sale which would not typically be included in the study of alcohol outlet density). We can see that there is some variation in accuracy between categories (see Fig. 1): bars, pubs and hotels have a higher representation than shops and restaurants (we only consider valid licenses when calculating these statistics), whilst the miscellaneous category is the least complete. Indeed, if we discard the miscellaneous category from consideration, the overall completeness

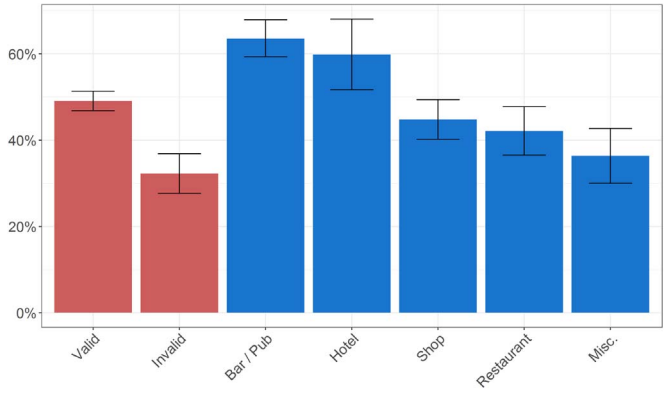

Fig. 1. Percentage of licenses found in OSM. Dataset is broken down into valid and invalid licenses (red bars) and by type of license (blue bars). Only valid licenses are considered for the license type statistics. Error bars represent 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of OSM data with respect to alcohol licenses rises to just over 52%.

It is also worth considering how much of the alcohol license data contained within the OSM database can be recovered by the automatic keyword searches we performed. Systematic construction of indicators of alcohol prevalence requires not only that venues licensed to sell alcohol are present within the OSM database, but also that they can be retrieved through keyword searches (i.e. that they have been correctly labelled as a bar, pub, shop, or other venue of interest). Hence we also looked at the volume of alcohol licenses which were both within the dataset and which could be discovered through the list of establishment types defined when we automatically retrieved data from OSM. We found that while valid alcohol point of sale data (excluding the miscellaneous license category) is around 52% complete in terms of records in OSM, the figure drops to 43% when we consider alcohol points of sale that could be retrieved with this type of search.

As a final part of our validation exercise, we seek to examine the extent to which OSM can be used to duplicate existing work on the relationship between alcohol availability and alcohol related harms. We tackle this by replicating the CRESH study which, as described above, looked at the relationship between alcohol availability and alcohol related harms in four cities in Scotland. We observe a strong positive correlation (R = 0.79) between our measure of alcohol outlet density and the measure reported by the CRESH study. The relatively high correlation between our measure and the CRESH study suggests that the use of the OSM generated measure should produce similar analytical results to the ones they generated. We formally test this by estimating similar regression models for on and off-site alcohol related harms to the ones found in their paper (Richardson et al., 2015, p176). The results of this process are shown in table one. As highlighted above, we cannot replicate their models exactly as we do not have the raw count data of hospitalisations and deaths for alcohol related harms which underpinned their model: instead, we have Standardised Mortality Rates for each Scottish Datazone in the four cities of interest. Hence, we produce ordered logistic regressions rather than Poisson count models.

As our modelling approach is different, in model 1 (Table 1) we simply replicate the results from the CRESH study by applying the CRESH data to this new modelling technique. Although the size of coefficients are difficult to compare, the direction and statistical significance of the results are the same, meaning that the use of a slightly different dataset and modelling approach does not affect the inferences drawn by Richardson et al. (2015), p176). Models 2–3 are identical to model 1 except they use estimates of alcohol density generated from OSM: model 2 looks at full alcohol density estimates based on OSM data; model 3 looks only at on-premises licenses and model 4 looks only at off-premises licenses. The conclusions are

Table 1
Replication of the CRESH study.

	Model 1: CRESH	Model 2: OSM Data	Model 3: Off-Premises	Model 4: On-Premises
CRESH density estimate	0.44***			
OSM density estimate		0.32***		
OSM: off-license only			0.36***	
OSM: on-license only				0.31***
Gender balance	− 0.09	− 0.17**	− 0.17**	− 0.17**
Average age	0.06	0.05	0.06	0.05
IMD Score	1.61***	1.57***	1.57***	1.57***
Aberdeen (reference)	1.00	1.00	1.00	1.00
Edinburgh	0.23	0.19	0.12	0.21
Dundee	0.59***	0.62***	0.68***	0.61***
Glasgow	0.64***	0.67***	0.68***	0.67***
Observations	1689	1689	1689	1689

Note: *p < 0.05.
** p < 0.01.
*** p < 0.001.

essentially identical to model 1: in all models, we also find a statistically significant positive correlation between density of alcohol outlets and incidence of alcohol related harms, with an effect size which is comparable (though smaller) to the one produced in the CRESH study. This supports the idea that OSM data can be used as a proxy for alcohol availability in statistical models analysing alcohol related harms.

2.2. Making use of OSM Quality Indicators

We have shown evidence above that OSM contains a considerable quantity of alcohol point of sale data, and can be used to duplicate existing results in alcohol research. However we have also shown that it is far from a complete record. The incompleteness in the data would naturally undermine confidence in any future work based on it. An important question is therefore if there would be any way of filtering out some of these imperfections, and hence selecting areas of only high quality OSM data to produce more trusted studies.

We investigate this possibility by making use of the OSM "quality indicators" described above, in two steps. First, we assign a quality rating to each of the postcode sectors which contained licenses in the validation dataset that was drawn from the LGA data. We then divide these sectors into quality deciles on the basis of the quality rating, and observe the strength of correlation between a quality rating and the completeness of the data found in our validation exercise. The results are shown in the left panel of Fig. 2. The bottom six deciles of English postcode sectors are broadly equivalent in terms of license completeness: all of them have between 40% and 45% of LGA licenses represented in their data. The seventh, eight and ninth deciles show an improvement to around 52%, while the tenth decile of the quality index has around 63% completeness.

In a second step, we assign an OSM quality rating to each Scottish Datazone, and then compare the correlation between the OSM data and the CRESH data in each decile (see the right panel of Fig. 2). The trend supports the results from England: in higher deciles of the quality index, the correlation between the OSM alcohol estimates and the CRESH measure is stronger, rising from just under 0.5 in the first decile to almost 0.85 in the tenth decile. Both of the above tests suggest that researchers could selectively identify areas of higher quality OSM data to perform analyses on. However, the data also shows that, even at the top of the quality index, the completeness of the OSM data is not perfect.

One potential objection to the strategy of selectively identifying

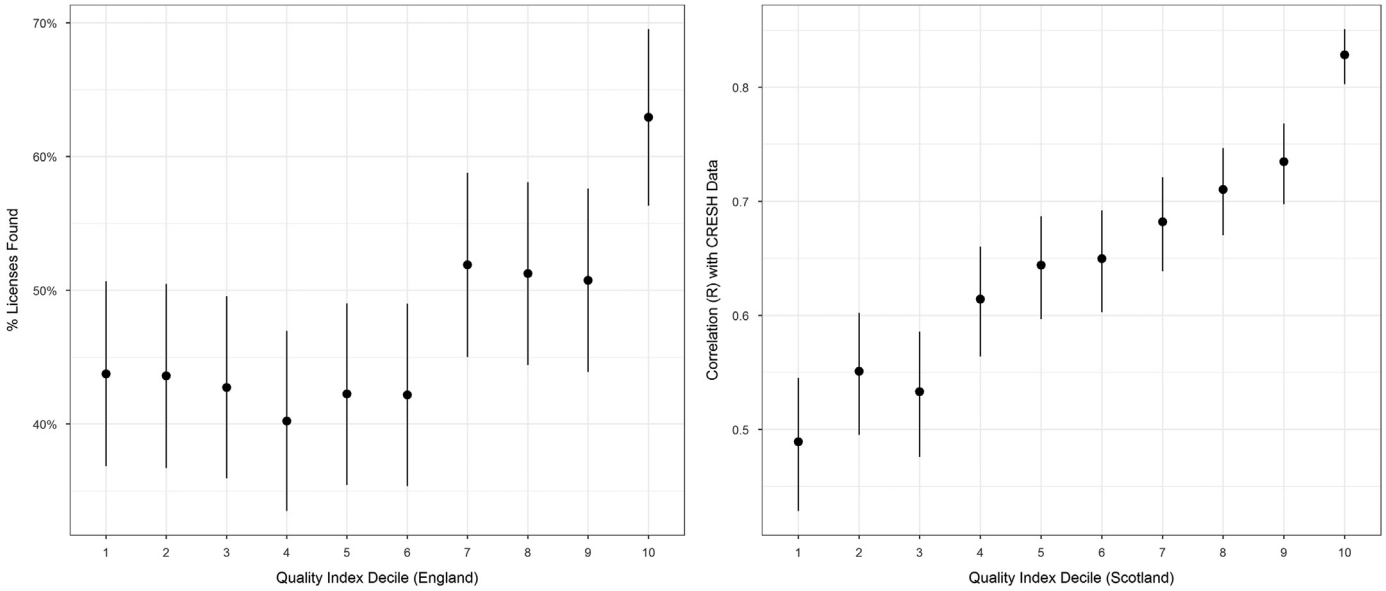

Fig. 2. Relationship between OSM Quality Indicators and alcohol completeness. Left panel shows how the percentage of licenses found in our validated LGA dataset varies at different deciles of the quality index; Right panel shows correlations between OSM and CRESH alcohol density estimates at different deciles of the quality index.

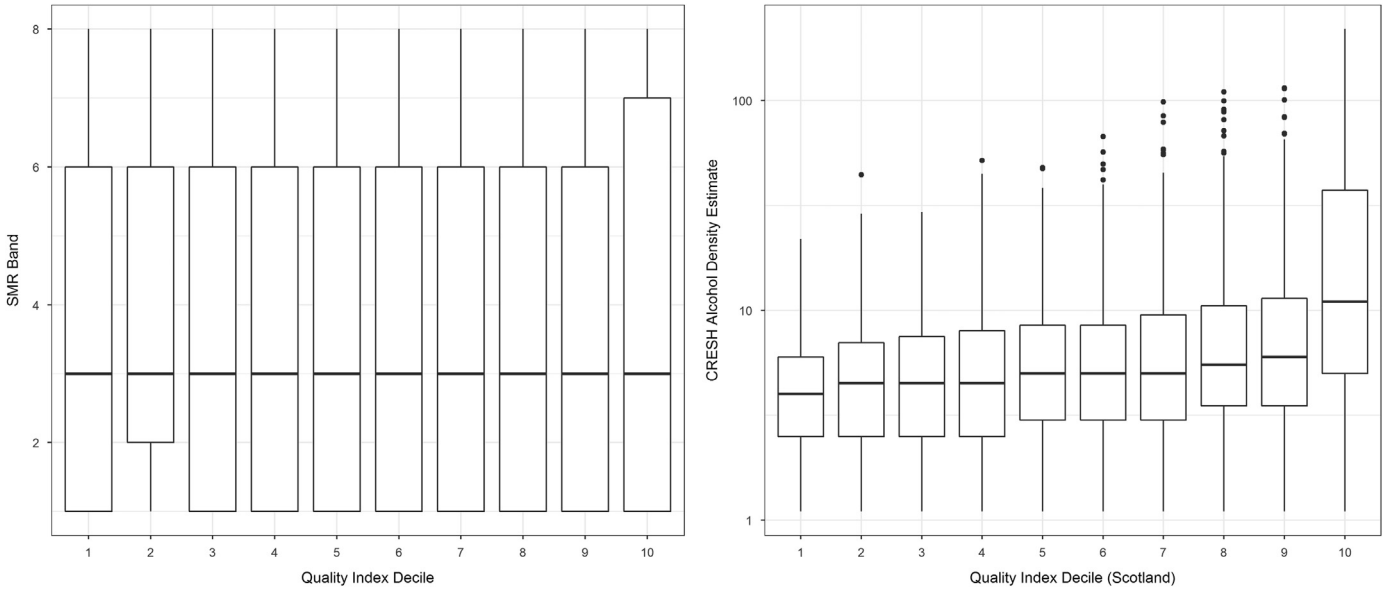

Fig. 3. Correlation between OSM quality index and both Standardised Mortality Rates (left panel) and CRESH Alcohol Density Estimates (right panel).

areas of higher quality data is that the distribution of OSM data quality may itself be correlated with factors of theoretical interest (either alcohol outlet density or alcohol related harms). Indeed, OSM data quality has been linked to socio-economic factors in the past (De Sabbata et al., 2016). Hence selecting high quality subsets might reduce variation in important dependent or independent variables. We investigate this possibility in Fig. 3, which looks at the relationship between the quality index of Scottish Datazones and both standardised mortality rates (left panel) and the CRESH alcohol density estimates (right panel). There is no observable relationship between the quality index and the standardised mortality rates: the boxplots are almost identical for each decile. There is by contrast a trend in terms of higher quality data being more likely to be found in areas with higher alcohol outlet density. However it is also the case that even in the upper deciles of the data, a significant range of alcohol density scores exists: the median alcohol outlet density is 3 in the first decile of the quality indicators, and still only 5 in the ninth decile (it rises to 10 in the tenth decile). Thus overall it seems that selectively making use of higher quality OSM data would not lead to great reductions in variation in key variables of theoretical interest.

3. Discussion and conclusion

The spatial availability of alcohol is considered to be a primary factor contributing to the prevalence and concentration of alcohol misuse and various physical and social harms (Babor et al., 2010; Stockwell and Gruenewald, 2003). While there is a wealth of evidence available to support this assertion (Campbell et al., 2009; Livingston et al., 2007; Popova et al., 2009), several recent studies have called for more rigorous and ambitious approaches (e.g. improved model specifications, greater context specificity and greater attention to theory-testing and refinement) to investigating the causal relationship between alcohol availability and its related health and social harms (Gmel et al., 2016; Holmes et al., 2014; Holmes and Meier, 2015). However, in the UK (as in many other countries) a major barrier preventing researchers and analysts further investigating this relationship has

been the inaccessibility and inconsistency of alcohol licensing data (Fry et al., 2016; Humphreys and Smith, 2013; Richardson et al., 2015). Although use of new data (e.g. market research) or innovations in data sharing, such as those introduced by the LGA, may lead to improvements in UK alcohol availability research, progress is likely to be slow without significant improvements in the availability of data to generate measures of spatial alcohol availability. In this study we have investigated an alternative approach: the potential use of volunteer geographic information on urban alcohol outlets available through OpenStreetMap (OSM).

The above validation exercise has shown promising evidence about the potential use of OSM data in alcohol research. Two contributions have been made in particular. Firstly, these analyses have shown that there is a significant quantity of alcohol license data freely available through OSM, and that estimates of spatial availability of alcohol generated with these data show good correlation with estimates generated from trusted external sources. Secondly, by incorporating indicators of OSM data quality, which can be generated from other data within the platform, we found that it is possible to predict where alcohol point of sale data within OSM would be more or less accurate, hence allowing researches to potentially select areas of higher data quality for studies (or at least measure the quality of the data they are working with). However, the study has also shown reasons to be cautious in the use of OSM data: for retail alcohol outlets we estimated that it is only around 52% complete (a figure which drops to 43% when we also consider how much data can be recovered through a keyword search), and when used in a replication of other analyses results are similar but not identical. These limitations make it clear that OSM data should not replace gold standard data collection efforts: to produce truly trusted conclusions, these higher quality data sources are necessary.

Despite these limitations, we do think that there is nevertheless a place for OSM data in the discipline. In particular, when researchers lack the financial resources to access trusted data sources, OSM can provide a useful supplement. We can envisage the data being used to quickly test ideas, and hence generate a proof of concept for further more involved research (for example, piloting studies which could then be supported by further data collection). We can also envisage it being used by PhD or early career researchers who lack the means to start major data collection efforts: in this respect it may provide a welcome spreading of opportunity within the field, and hence allow a wide variety of new theories to be tested, albeit in an exploratory fashion. Indeed, as we have described above, volunteer geographic information is being applied to a wide range of public health research domains, and our study provides suggestive evidence that it may also be useful in other areas of work (for example, mapping the spatial availability of food, or green spaces, or other public amenities), though further specific validation exercises would be needed to confirm this. In short, cautious use of OSM data, with appropriate caveats, could help advance our understanding of the relationship between alcohol availability and a variety of health and social harms.

Acknowledgements

This research was partially funded by a grant from the ESRC (Grant no. ES/M010058/1).

Appendix A

Selecting Data from OpenStreetMap

The OpenStreetMap database contains records on all of the features found on OpenStreetMap: from natural features such as roads, rivers and hills to human created ones such as parks, schools, hospitals, shops and restaurants. Each data record in the database also has a variety of "tags" attached to it which help describe what the record relates to (a full list of these tags can be found here: http://wiki.openstreetmap.org/wiki/Map_Features). We selected all data which had at least one of the following tags: "building", "shop", or "amenity". Each tag can also take a variety of values, indicating what in particular it refers to. For example, a map feature could be tagged with "shop = alcohol" or "amenity = pub". We created a list of tag – value pairs which we believed were likely to indicate venues which sold alcohol. The complete list of these pairs can be found in Table A1 below.

Any records in the OSM database which had at least one of the tag – value pairs described in Table A1 were selected into our data, and hence used to generate the alcohol density estimates.

Creating OSM quality indicators for Scottish Datazones

The quality of data in OSM can vary according to geographic location: it depends, essentially, on how many users are willing to contribute in a given area, and how many contributions they make. As contribution levels can themselves be measured from data freely released by OSM, it is therefore possible to create metrics which measure the likely quality of data in an area, given patterns of use by individuals in the area.

The quality metric we make use of in this paper is composed of a variety of separate indicators of the extent of user contribution in a given area (see Table A2). Their selection was inspired by the work of Barron et al. (2014) who proposed some general principles for measuring the quality of data in a given area. The quality metric itself is simply an average of the value of each indicator. Before averaging them together, some of the indicators are transformed (so their distributions become more normal); all indicators are also standardised, so that each one makes the same average contribution to the index (the transformation applied to each indicator is also listed in the table).

Table A1
List of tags and values used for the OSM Database Query.

Tag name	List of values
Amenity	Bar, Pub, Restaurant, Biergarten, Café, Fast Food, Cinema, Nightclub, Theatre
Shop	Alcohol, Beverages, Brewing Supplies, Convenience, Wine, General, Supermarket, Newsagent
Building	Hotel

Table A2
List of map features used in the construction of the quality index.

OSM area usage indicator	Transformation
Number of map features per square kilometre	Log transformed and then scaled
Average number of edits per map feature	Log transformed and then scaled
Number of users who made at least one edit in the area	Log transformed and then scaled
Number of features edited by more than one user	Log transformed and then scaled
Number of days with at least one edit in the area	Log transformed and then scaled
Days since the last edit in the area	Inversed and then scaled
Proportion of buildings with full address information	-1 if 0%, set to 0 if no buildings in area

References

Angus, C., Holmes, J., Maheswaran, R., Green, M.A., Meier, P., Brennan, A., 2017. Mapping patterns and trends in the spatial availability of alcohol using low-level geographic data: a case study in England 2003–2013. Int. J. Environ. Res. Public Health 14 (4), 406. http://dx.doi.org/10.3390/ijerph14040406.

Arsanjani, J.J., Mooney, P., Zipf, A., Schauss, A., 2015. Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In: Arsanjani, J., Zipf, A., Mooney, P., Helbich, M. (Eds.), OpenStreetMap in GIScience. Springer International Publishing, Cham, 37–58.

Babor, T., Caetano, R., Casswell, S., Edwards, G., Giesbrecht, N., Graham, K., Rossow, I., 2010. Alcohol: No Ordinary Commodity, Research and Public Policy 2nd ed.. Oxford University Press, Oxford.

Barron, C., Neis, P., Zipf, A., 2014. A comprehensive framework for intrinsic OpenStreetMap quality analysis. Trans. GIS 18 (6), 877–895. http://dx.doi.org/10.1111/tgis.12073.

Bergquist, R., Rinaldi, L., 2010. Health research based on geospatial tools: a timely approach in a changing environment. J. Helminthol. 84 (1), 1–11.

Boulos, M.N.K., Resch, B., Crowley, D.N., Breslin, J.G., Sohn, G., Burtner, R., Pike, W.A., Jezierski, E., Chuang, K.-Y., 2011. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. Int. J. Health Geogr. 10, 67.

Bright, J., De Sabbata, S., Lee, S., 2017. Geodemographic biases in crowdsourced knowledge websites: do neighbours fill in the blanks? Geojournal. http://dx.doi.org/10.1007/s10708-017-9778-7.

Bryden, A., Roberts, B., McKee, M., Petticrew, M., 2012. A systematic review of the influence on alcohol use of community level availability and marketing of alcohol. Health Place 18 (2), 349–357. http://dx.doi.org/10.1016/j.healthplace.2011.11.003.

Campbell, C.A., Hahn, R.A., Elder, R., Brewer, R., Chattopadhyay, S., Fielding, J., Middleton, J.C., 2009. The effectiveness of limiting alcohol outlet density as a means of reducing excessive alcohol consumption and alcohol-related harms. Am. J. Prev. Med. 37 (6), 556–569. http://dx.doi.org/10.1016/j.amepre.2009.09.028.

Cinnamon, J., Schuurman, N., 2013. Confronting the data-divide in a time of spatial turns and volunteered geographic information. GeoJournal 78 (4), 657–674.

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. Montreal, Canada.

Fone, D., Morgan, J., Fry, R., Rodgers, S., Orford, S., Farewell, D., Lyons, R., 2016. Change in Alcohol Outlet Density and Alcohol-related Harm to Population Health (CHALICE): A Comprehensive Record-linked Database Study in Wales. NIHR Journals Library, Southampton, UK⟨http://www.ncbi.nlm.nih.gov/books/NBK350758/⟩, (Last accessed 15 June 2017).

Fry, R.J., Rodgers, S.E., Morgan, J., Orford, S., Fone, D.L., 2016. Using routinely collected administrative data in public health research: geocoding alcohol OutletData. Appl. Spat. Anal. Policy, 1–15. http://dx.doi.org/10.1007/s12061-016-9184-4.

Girres, J.F., Touya, G., 2010. Quality assessment of the French OpenStreetMap dataset. Trans. GIS 14 (4), 435–459. http://dx.doi.org/10.1111/j.1467-9671.2010.01203.x.

Gmel, G., Holmes, J., Studer, J., 2016. Are alcohol outlet densities strongly associated with alcohol-related outcomes? A critical review of recent evidence. Drug Alcohol Rev. 35 (1), 40–54. http://dx.doi.org/10.1111/dar.12304.

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

Goodchild, M.F., Li, L., 2012. Assuring the quality of volunteered geographic information. Spat. Stat. 1, 110–120.

Goranson, C., Thihalolipavan, S., & di Tada, N., 2012. VGI and Public Health: Possibilities and Pitfalls. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice, 329–340. https://doi.org/10.1007/978-94-007-4587-2_18.

Hadfield, P., 2006. Bar Wars: Contesting the Night in Contemporary British Cities. Oxford University Press, Oxford.

Haklay, M., 2010a. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environ. Plan. B: Plan. Des. 37 (4), 682–703. http://dx.doi.org/10.1068/b35097.

Helbich, M., Amelunxen, C., Neis, P., Zipf, A., 2012. Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata. In: Jekel, T., Car, A., Strobl, J., Griesebner, G. (Eds.), Geospatial Crossroads @ GI_Forum '12. Proceedings of the Geoinformatics Forum Salzburg, 24–33.

HM Government, 2012. The Government's Alcohol Strategy. HM Governmen, London.

HMSO, 2005. The Licensing Act 2003, Chapter 17. Available at: ⟨http://www.legislation.gov.uk/ukpga/2003/17⟩ (Last accessed 15 June 2017).

Holmes, J., Guo, Y., Maheswaran, R., Nicholls, J., Meier, P.S., Brennan, A., 2014. The impact of spatial and temporal availability of alcohol on its consumption and related harms: a critical review in the context of UK licensing policies. Drug Alcohol Rev. 33 (5), 515–525. http://dx.doi.org/10.1111/dar.12191.

Holmes, J., Meier, P.S., 2015. Commentary on Hobday et al. (2015): inconsistent results beneath consistent conclusions—the need for a new approach to analysing alcohol availability. Addiction 110 (12), 1910–1911. http://dx.doi.org/10.1111/add.13180.

Humphreys, D.K., Eisner, M.P., 2010. Evaluating a natural experiment in alcohol policy: the licensing act (2003) and the requirement for attention to implementation. Criminol. Public Policy 9 (1), 41–67. http://dx.doi.org/10.1111/j.1745-9133.2010.00609.x.

Humphreys, D.K., Eisner, M.P., 2014. Do flexible alcohol trading hours reduce violence? A theory-based natural experiment in alcohol policy. Soc. Sci. Med. 102, 1–9.

Humphreys, D.K., Smith, D.M., 2013. Alcohol licensing data: Why is it an underused resource in public health? Health Place 24, 110–114. http://dx.doi.org/10.1016/j.healthplace.2013.07.006.

Livingston, M., Chikritzhs, T., Room, R., 2007. Changing the density of alcohol outlets to reduce alcohol-related problems. Drug Alcohol Rev. 26 (5), 557–566. http://dx.doi.org/10.1080/09595230701499191.

Mashhadi, A., Quattrone, G., Capra, L., 2015. The impact of society on volunteered geographic information. In: Arsanjani, J., Zipf, A., Mooney, P., Helbich, M. (Eds.), OpenStreetMap in GIScience. Springer International Publishing, Cham, 125–141.

Meenar, M., 2017. Using participatory and mixed-methods approaches in GIS to develop a Place-Based Food Insecurity and Vulnerability Index. Environ. Plan. A.

Mooney, P., Corcoran, P., 2011. Integrating volunteered geographic information into pervasive health computing applications. In: Proceedings of the 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops.

Pfeiffer, D., Stevens, K.B., 2015. Spatial and temporal epidemiological analysis in the Big Data era. Prev. Vet. Med. 122 (1–2), 213–220.

Popova, S., Giesbrecht, N., Bekmuradov, D., Patra, J., 2009. Hours and days of sale and density of alcohol outlets: impacts on alcohol consumption and damage: a systematic review. Alcohol Alcohol. 44 (5), 500–516. http://dx.doi.org/10.1093/alcalc/agp054.

Quinn, S., Yapa, L., 2015. OpenStreetMap and food security: a case study in the City of Philadelphia. Prof. Geogr., 1–10.

Richardson, E.A., Hill, S.E., Mitchell, R., Pearce, J., Shortt, N.K., 2015. Is local alcohol outlet density related to alcohol-related morbidity and mortality in Scottish cities? Health Place 33, 172–180. http://dx.doi.org/10.1016/j.healthplace.2015.02.014.

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. Int. J. Geogr. Inf. Sci. 31 (1), 139–167. http://dx.doi.org/10.1080/13658816.2016.1189556.

Stensgaard, A.-S., Saarnak, C., Utzinger, J., Vounatsou, P., Simoonga, C., Mushinge, G., Rahbek, C., Møhlenberg, F., Kristensen, T.K., 2009. Virtual globes and geospatial health: the potential of new tools in the management and control of vector-borne diseases. Geospatial Health 3 (2), 127–141.

Stockwell, T., Gruenewald, P., 2003. Controls on the Physical Availability of Alcohol. In: Heather, N., Stockwell, T. (Eds.), The Essential Handbook of Treatment and Prevention of Alcohol Problems. John Wiley & Sons Ltd, Chichester, 213–234.

Zielstra, D., Zipf, A., 2010. May. A comparative study of proprietary geodata and volunteered geographic information for Germany. In: Proceedings of the 13th AGILE international conference on geographic information science (Vol. 2010).