

Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <http://orca.cf.ac.uk/88523/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Fry, Richard J., Rodgers, Sarah E., Morgan, Jennifer, Orford, Scott and Fone, David Lawrence 2017. Using routinely collected administrative data in public health research: geocoding alcohol outlet data. *Applied Spatial Analysis and Policy* 10 (2) , pp. 301-315. 10.1007/s12061-016-9184-4 file

Publishers page: <http://dx.doi.org/10.1007/s12061-016-9184-4> <<http://dx.doi.org/10.1007/s12061-016-9184-4>>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Using Routinely Collected Administrative Data in Public Health Research: Geocoding Alcohol Outlet Data

Richard J. Fry¹ · Sarah E. Rodgers¹ ·
Jennifer Morgan² · Scott Orford³ · David L. Fone²

Received: 20 May 2015 / Accepted: 4 March 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract We describe our process of geocoding alcohol outlets to create a national longitudinal exposure dataset for Wales, United Kingdom from 2006 to 2011. We investigated variation in the availability of data items and the quality of alcohol outlet addresses held within unitary authorities. We used a standard geocoding method augmented with a manual matching procedure to achieve a fully spatially referenced dataset. We found higher quality addresses are held for outlets based in urban areas, resulting in the automatic geocoding of 68 % of urban outlets, compared to 48 % in rural areas. Missing postcodes and a lack of address structure contributed to a lower geocoding proportion. An urban rural bias was removed with the development of a manual matching procedure. Only one-half of the unitary authorities provided data on on/off sales and opening times, which are important availability factors. The resulting outlet dataset is suitable for contributing to the evidence-base of alcohol availability and alcohol-related harm. Local government should be encouraged to use standardised data fields, including addresses, to enable accurate geocoding of alcohol outlets and facilitate research that aims to prevent alcohol-related harm. Standardising data collection would enable efficient secondary data reuse using record linkage techniques, allowing

✉ Sarah E. Rodgers
s.e.rodgers@swansea.ac.uk

¹ Farr Institute, Swansea University Medical School, Singleton Campus, Swansea SA2 8PP, UK

² Institute of Primary Care & Public Health, Cardiff University School of Medicine, Neuadd Meirionnydd, Heath Park, Cardiff CF14 4YS, UK

³ School of Planning and Geography, Cardiff University, Glamorgan Building, King Edward VII Avenue, Cardiff, Wales CF10 3WA, UK

the retrospective creation and evaluation of population-based natural experiments to provide evidence for policy and practice.

Keywords Geocoding · Administrative data · Alcohol outlets · Alcohol-related harm · Natural experiment

Introduction

The ‘natural experiment’ is an important methodology for investigating the effectiveness of public health interventions (Cousens et al. 2009; Craig et al. 2012; Ogilvie et al. 2006; Petticrew et al. 2005). Natural experiments can be the best source of available evidence where the randomised controlled trial is not feasible and the intervention(s) have a natural non-random variability in geographical location and time. Many public health interventions, such as urban regeneration schemes, (Egan et al. 2010; Petticrew et al. 2009) are based in small geographical areas and so the spatial scale of measurement of exposures and outcomes that change over time becomes an important consideration.

Methods to evaluate the impact of natural experiments have been enhanced by the development of anonymised electronic record-linkage in secure databanks that facilitate the use of administrative data in public health research (Fone et al. 2012, 2013; Rodgers et al. 2012; White et al. 2014). Data linkage methods allow spatially referenced exposures to be linked to existing cohort studies and to populate age-sex registers for defining new cohorts with full information on small-area migration. These new rich sources of data are facilitating a new approach to the longitudinal analysis of natural experiments (Fone et al. 2013; Rodgers et al. 2012; White et al. 2014).

The main challenge for the use of routine administrative data in measuring variation in exposure across space and time is fulfilling the requirement for good quality standardised data. Ideally, the time and precise geographical location of unplanned changes that have taken place, due to policy implementation or other factors, should be recorded with high resolution to enable the longitudinal variation in exposure to be used in a natural experiment. Without this, spatial measures of exposure will be biased from misclassification of events across spatial boundaries and from events that are missing geographical or temporal identifiers.

We have recently carried out a natural experiment of the effect of change in alcohol availability on alcohol-related harm in the community (CHALICE) (Fone et al. 2012). Excess alcohol consumption is known to cause a wide range of harm, including liver disease (Leon and McCambridge 2006), cancers (Parkin 2011), cardiovascular disease (Klatsky and Gunderson 2008; Reynolds et al. 2003), suicide (Ramstedt 2001), violence and injuries (Flatley et al. 2010; Popova et al. 2009). One method proposed to reduce the societal burden of alcohol-related harm is to reduce the availability of alcohol through a reduction in the density of alcohol outlets (Babor 2008; British Medical British Medical 2008; Humphreys and Smith 2013; Young et al. 2013; Gruenewald 2011).

To investigate this further we had to define and measure outlet density. The problem that we address in this paper is the first step in estimating an outlet density, namely how to identify, enumerate and locate alcohol outlets in geographical space. This required the use of spatially referenced administrative data to quantify alcohol availability. Three Scottish studies have previously used local licensing authority data on alcohol outlets.

Two studies set in Glasgow found that the unit postcode was obtained for every outlet recorded in 2006, but did not report any method of validation (Ellaway et al. 2010; R. Young et al. 2013). A more recent study quantified the number of alcohol outlets for each of the 6505 small-area datazones in Scotland provided by local Licensing Boards in 2012 (Richardson et al. 2014). After dataset cleaning and removal of duplicates they recorded 222 (1.4 %) fewer outlets than the official Scottish Liquor Licensing Statistics. However the variation between the 32 Licensing Boards in Scotland was from 7.4 % fewer to 1.8 % more outlets and this degree of variation will be more problematic for preparing accurate counts at smaller geographical levels.

We report on aspects of the CHALICE project (Fone et al. 2012) where we collated, geocoded address-level data, and analysed 3-monthly changes in the number of alcohol outlets in Census Lower Layer Super Output Areas across Wales for use in a longitudinal analysis of the impact of change in alcohol outlet density on alcohol-related harm. We discuss the challenges faced in the secondary use of publicly available alcohol licence data for household and small-area research and describe the methods we developed to geocode the licence data in order to estimate outlet density in a Geographic Information System. Humphreys and Smith (2013) document the difficulties in collecting data from unitary authorities. Our study aims to advance the discussion by assessing the technical challenges researchers face following receipt of the data in achieving a dataset ready for analysis.

Study Design

The setting is Wales, a country in the UK with a population of just over 3 million. More than one-half of the land area is rural with 11 % of the population living in the sparsest rural category; almost 2 million people live in urban areas (Pateman 2011). Local government is organised into 22 unitary authorities that oversee many administrative functions and as such are an important source of routinely collected data. The study area is illustrated in Fig. 1.

Licence Data Acquisition

An alcohol outlet refers to any premise or club licensed by a unitary authority to sell alcohol. The 22 Welsh unitary authorities are responsible under The Licensing Act (2003), for maintaining public registers of all licensed alcohol outlets (Hadfield et al. 2010). The Act came into force in England and Wales on 24th November 2005. Licences may be in the form of a premise licence, a club licence or a temporary event notice. The Act requires licensing sub-committees to record the name and address of each outlet, including postcode details, using a licensing pro forma. A completed pro forma is then submitted to the unitary authority for consideration; successful applications are recorded and made available for public inspection.

The use of unitary authority data offers a potentially unbiased data source with no marketing or competitive agenda. We requested the names and location of all licensed alcohol outlets from each unitary authority for dates between Nov 2005 and Dec 2011, using a similar model to the one described by Humphreys and Smith (2013). The data items requested for each outlet were: 1) the date permission was granted or the licence became active (referred to as *Start Date*), 2) the licence expiry date or an indicated date of

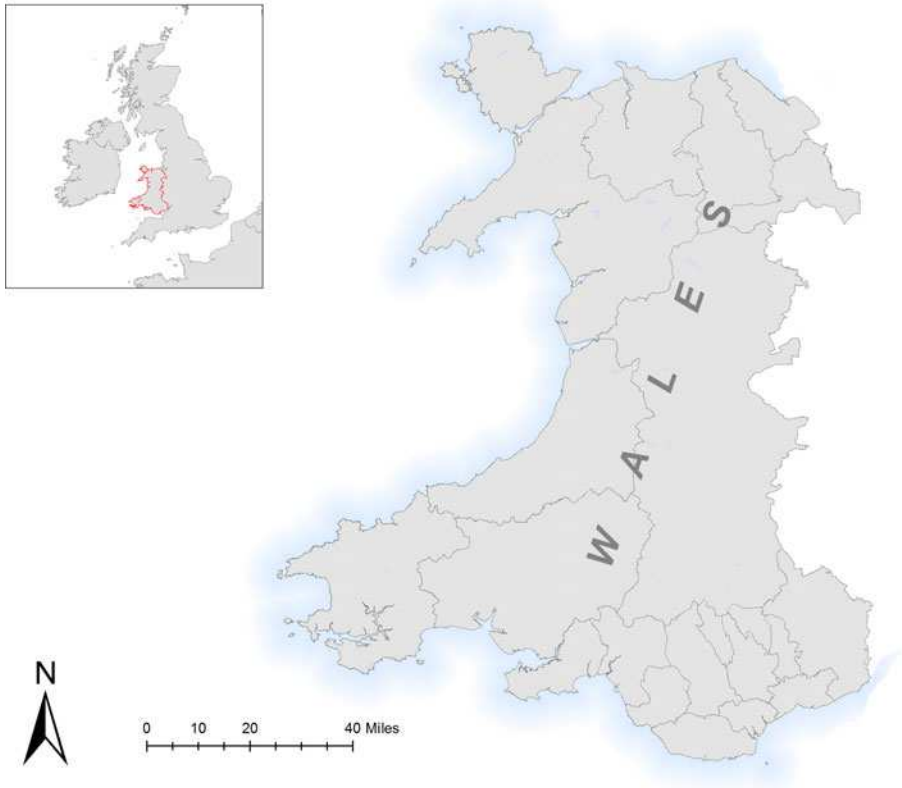


Fig. 1 Study Area, with unitary authority boundaries

outlet closure (*End Date*), 3) whether the premise is licensed for ON and/or OFF premise sales (*On/Off Status*), 4) the hours permissible to sell alcohol or general opening hours of the outlet (*Opening Hours*), and 5) the type of premise as assigned by the unitary authority (*Premise Type*), if available. Under The Act, unitary authorities are mandated to maintain public registers of current licences, populated using licence application forms, which include items 1, 3 and 4 only. The registers hold data about premises that have licences to cover food and music; therefore we selected only those relating to alcohol.

Licence Data Format

Consistent with the findings of Humphreys and Smith (2013), outlet data were received in a wide variety of media and with different amounts of detail. Only one-half of the unitary authorities provided data on the on/off premise sales and opening times. A similar number provided data in a spreadsheet; several provided pdfs or electronic text files, and one provided printed copy only. Historical outlet data from January 2006 onwards were received from 21 of the 22 authorities. One authority could provide only their current (2012) public register. An actual or approximate licence start date was used to summarise the number of outlets into time periods for 21 of the 22 authorities. Several were unable to provide precise outlet closure dates, so the date of the last

interaction with the outlet by the authority was used to generate an approximate end date. Each authority presented its own unique challenges for data gathering, such as the need to collect records in person; collation from multiple pdfs; text files without consistent formatting; or print-outs that needed to be converted into a useable format. As a result, extensive processing was required to gather all files, transpose the information into useable outlet records, and collate into a single database of outlets. The tools and techniques used included web scraping, string manipulation and optical character recognition. The final database consisted of 21,137 outlets. Each record contained some address information (partial or full) for the premise; we describe in the next section how we used the available address details to geocode outlets.

Methods

Geocoding Licence Data

Geocoding is the process of matching text-based address data to known geographic coordinates. There have been a number of studies that have assessed geocoding techniques, positional accuracy (Zandbergen 2008; Ward et al. 2005; Duncan et al. 2011) and the impacts of geocoded or matched outlet proportion and positional accuracy upon resulting analyses (Krieger et al. 2001; Ratcliffe 2004; Zandbergen 2007). In particular, Zandbergen (2007) has shown that wherever possible, high resolution address data (building-level) should be employed as the reference dataset particularly when fine-scale analysis is being performed. Given that the CHALICE project looks at the alcohol availability for a 10-minute walking and driving buffer around a residence, it was important to use the best quality address data available for the study area.

Ordnance Survey AddressBase® Premium (Ordnance Survey 2014) was used as the reference dataset for the geocoding procedure. AddressBase® Premium is the most comprehensive address dataset for Great Britain and is based upon the Royal Mail Postcode Address File (PAF) and National Land and Property Gazetteer (NLPG). The AddressBase® Premium dataset contains a point for each residence, which is located within the footprint of the residence. The buildings are surveyed with a spatial accuracy of $\pm 1\text{--}2$ m in urban areas. Although this dataset is comprehensive and spatially accurate, it was not suitable as our source of outlet dataset because it did not contain a suitable classification to extract all alcohol outlets, neither did it contain information specific to alcohol outlets such as on/off sales or opening times.

A geocoding script creating an ArcMap plugin was written in VB.NET¹ to break down each outlet address received from the unitary authorities into its component parts (Organisation Name, Building Name, Street Number, Thoroughfare, Post town and Postcode) as defined by the Royal Mail (Royal 2010). Each address component was indexed before the reference AddressBase® Premium dataset was searched and the best match was recorded. The best match was achieved using string matching and regular expression techniques to resolve any formatting/grammatical issues (e.g. commas, hyphens, capitalisation). For records that were not fully matched in a first pass, a probabilistic matching algorithm was implemented to identify addresses based on

¹ Available on request

partial address matches. For example, a premise was identified from a street number and name where a premise name could not be matched. The final geocoding step filtered for matches based on the postcode provided in the licence data.

Each successfully fully matched unitary authority outlet address record was allocated the corresponding Unique Property Reference Number (UPRN) from the AddressBase® Premium data, as well as the relevant geographic coordinates. Only at this stage was the dataset deemed to be ready for analysis, specifically suitable for utilisation in longitudinal outlet density research.

Manually Matching Licence Data

Outlets that were not geocoded underwent a manual matching process. There is a precedent for using web-based mapping technologies to fill gaps and perform neighbourhood audits (Clarke et al. 2010; Rundle et al. 2011; Rossen et al. 2012). Google Maps and Google StreetView were examined for the outlet location using the information contained in the licence record, including, for example, the outlet name, street and locality. The outlet was identified as a point on Google Maps and the latitude and longitude of the outlet location were extracted using the information for the point contained in the URL. For any remaining outlets, for which the exact outlet location was unclear, further effort was made to approximate the location within the appropriate street by assessing the function of buildings using Google Streetview. This third “approximated” stage assigned all outlet locations in order to achieve a final dataset that was suitable for spatio-temporal research to guide policy and practice.

A validation exercise of the manual matching was completed using a 10 % random sample of geocoded outlets. This resulted in a dataset with two sets of coordinates; the geocoded outlet location, treated as the “true location”, and the manually matched location. Location errors, introduced as part of the manual match process, were calculated by measuring the Euclidean distance between the two sets of coordinates. The final dataset has the match type (geocoded, manually matched, and approximated) recorded against each outlet. This will enable purely spatial future research focussed on urban areas to omit the approximated locations.

The geocoded outlets will be used in future work to calculate alcohol density scores using a GIS. To investigate whether there would be large differences in the proportion of outlets matched between urban and rural authorities, we classified the outlets into one of six rural–urban classes of settlement type based on the ONS rural urban classification (ONS Geography 2004) assigned to the small area geographies (Lower Super Output Area). For each unitary authority the number of urbanised LSOAs were aggregated into percent urbanised.

Results

Geocoding Licence Data

Of the 21 137 individual licence records received, 16 101 remained after filtering out those that were not alcohol licences, were temporary event notices, or duplicates. Of

these, 8536 (51 %) were matched using geocoding software (matched against Ordnance Survey AddressBase® Premium). Temporal coverage of outlets was calculated across all unitary authorities for 6 years (January 2006 to December 2011) apart from one authority that retained only current licence records.

Successful geocoding proportions varied from 28 to 72 % (Table 1). The geocoded proportion was influenced by the structure and completeness of unitary authority addresses (data not shown). All provided records contained some address information but several unitary authorities only provided an address in a single column. Some outlet records contained addresses with a missing postcode field suggesting address quality is reduced, although it may be an optimistic measure because some of the provided postcodes may be incorrect.

The main reasons geocoding failed were an absent or inaccurate premise name or street name, or an absent or inaccurate postcode. For example, a relatively small change from number 10 *Market Square*, to 10 *Market Street*, and other small errors in the main address elements made it difficult for a machine to decide between multiple possible matches. This was particularly true when the error in the address partially matched an alternative address. The example outlined previously illustrates this because both Market Square and Market Street exist in the locality, making identifying the correct address impossible for an automated system.

There was no pattern of geocoding success by the proportion of LSOAs in the most deprived 10 % of the Welsh Index of Multiple Deprivation (Local Government Data Unit - Wales 2005) (data not shown), nor was there a historical influence. The average geocoding proportion was 51 % for all years, ranging from 53 % in 2006 to 51 % in 2011. The geocoded proportion, however, was generally higher for records obtained from more urbanised unitary authorities compared to those authorities in a more rural setting (Fig. 2). The percentage of records successfully matched is shown in Fig. 2 for each ONS urban–rural class. The unitary authorities were arranged in order of percent urbanised. Outlet records from three cities in south Wales; Cardiff, Swansea and Newport, had geocoding proportions of 68, 64 and 65 %, respectively, averaging 66 %. In contrast, rural geocoding proportions ranged from 28 to 68, averaging 48 %.

Manually Matching Licence Data

Records that were not geocoded underwent a manual matching procedure using Google Streetview to identify an outlet location. The proportion of records manually matched and those approximated are shown in Table 1. The manual matching validation process for a 10 % sample (1604) of geocoded outlets, revealed that 79 % were manually matched to locations within 100 m Euclidean distance of the AddressBase® Premium location, 90 % within 500 m and 93 % within 1000 m. For reporting purposes data were aggregated into standard census geographies (LSOA). Several geocoded outlets (<8 %) were located in a different LSOA when manually matched but the spatial distribution revealed no systematic error (Fig. 3). The validation process was completed to give confidence in the accuracy of the combined geocoding and manual matching procedures in allocating an accurate spatial location to alcohol outlets, and assure the usefulness of these outlet data for aggregation into areas defined by travel distances from each household.

Table 1 Match proportion summary for the alcohol outlets against Ordnance Survey AddressBase® Premium for each processing stage

Unitary authority	Geocoded (count)	(%)	Manually matched (count)	(%)	Approximated (count)	(%)	% Missing opening hours	% Missing On/OFF Status
1	94	32	203	68	0	0	100	88
2	251	51	242	49	2	0	100	9
3	409	72	147	26	11	2	100	100
4	901	68	416	31	5	0	0	0
5	614	68	280	31	11	1	100	0
6	290	58	165	33	41	8	100	0
7	339	49	292	42	58	8	59	10
8	288	49	248	42	52	9	1	0
9	500	35	876	61	55	4	1	0
10	451	56	303	38	50	6	3	0
11	140	38	224	60	9	2	100	100
12	136	56	103	42	6	3	100	100
13	130	28	325	70	12	3	7	18
14	263	50	244	46	24	5	20	99
15	1035	64	557	34	22	1	100	0
16	431	45	486	50	45	5	3	15
17	438	43	528	51	61	6	100	100
18	378	58	271	42	4	0	0	0
19	608	65	322	34	9	1	100	0
20	432	55	351	45	3	0	100	0
21	199	43	246	53	17	4	100	2
22	209	47	220	49	19	4	100	100
Total / Average	8536	51	7049	45	516	3	37	33

Averages calculated from data prior to rounding

Discussion

We found that current and historical alcohol outlet licence data are collected and retained within most unitary authorities in Wales. There were, however, important differences between the authorities in obtaining access to the licensing data, and substantial variation in the data items stored. For example, information on premise type for on- and off-sales and opening hours were recorded by only one-half of the unitary authorities. There were substantial differences in the quality and accuracy of the address details captured, which impacted on the geocoding process and required the introduction of a three-stage matching process (geocoded, manually matched and approximated). This was necessary to achieve the most accurate count of outlets at household- and LSOA-level for each three month period in the study, particularly important for accurately estimating quarterly change in outlet density.

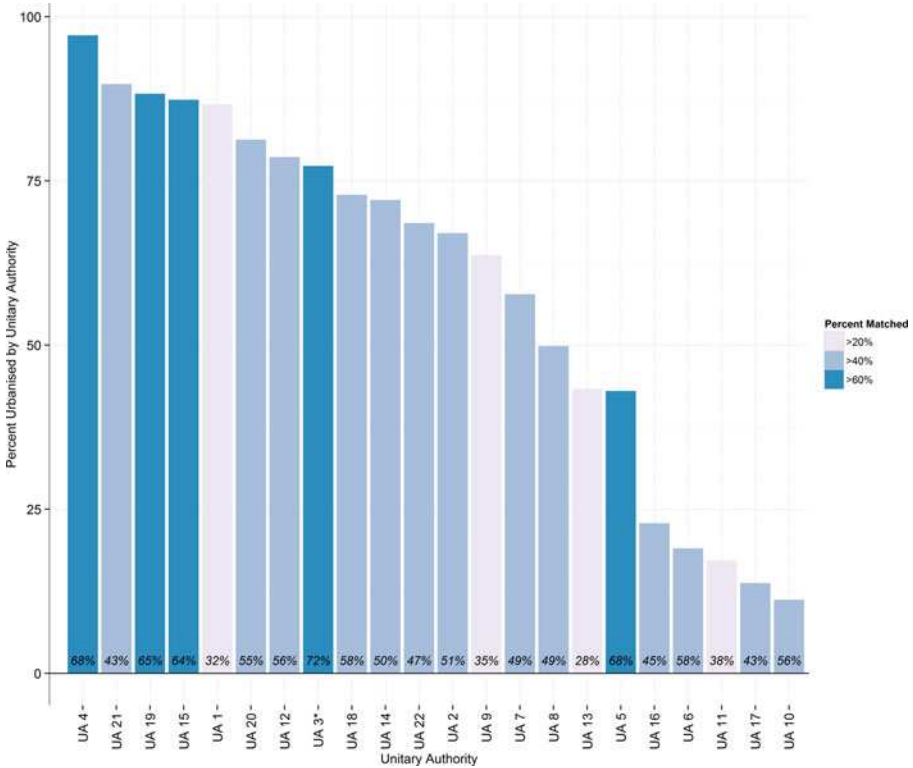


Fig. 2 Comparison of urban percent classification, derived from ONS rural–urban classification (*left axis*), and geocoded percentage (legend). *Proportion matched for UA3 for 1 year due to receipt of current outlets only

As Humphreys and Smith (2013) found, the effort required to obtain a dataset ready for analysis was considerable including problems in accessing and preparing these data for use. We initially under-estimated the length of time required to obtain the Licensing Act data. Several follow-up requests had to be made to ensure historical, rather than only current, licensing data were provided, with many unitary authorities unclear how to obtain such information from their systems. Six unitary authorities required as many as ten follow-up calls, and up to three months of data checking time before release. In some cases Freedom of Information (FOI) requests were necessary. One unitary authority deleted old records and therefore could not provide historical data. The availability of historical data in all but one unitary authority was nevertheless an improvement over other UK datasets, approaching the quality of a researcher-led dataset held for New Orleans (Xu et al. 2012) and data that are readily available in Finland (Halonen et al. 2013).

The bilingual status of Wales had an impact on geocoding. A licence application can be made in Welsh or English, and records may be administered and maintained in either language. This leads to different formats and spellings in licence records. This is particularly true of address fields, which contained misspelled outlet names and streets, resulting in no matches found during the geocoding process. It also makes the manual matching process more challenging.

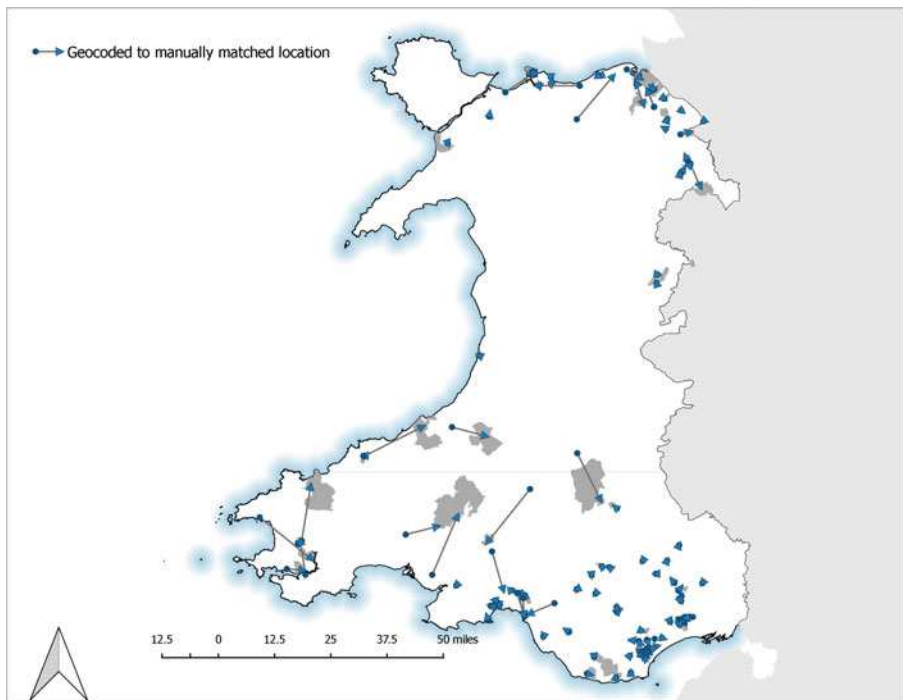


Fig. 3 Map illustrating outlet location error as a result of the manual matching process for a validated sample of 1604 geocoded outlets

We achieved a relatively low overall geocoded proportion of 51 %. To improve this we developed a manual matching method to locate premises that could not be geocoded. The AddressBase[®] Premium dataset, used as the geocoding data source, is in part generated and maintained by the same unitary authorities approving licence applications. From this low resulting geocoding proportion, we concluded that basic checks on vital address components, such as postcodes or building names and numbers, had not taken place. These are essential to achieve accurate and fast geocoding. We briefly considered using Points of interest (POI) data instead of those sourced from unitary authorities but the POI dataset does not include historical outlet data to complete longitudinal research and, additionally, previous researchers found that the POI dataset contained fewer than 64 % of the outlets known to unitary authorities (Burgoiné and Harrison 2013). This degree of error is too high for spatio-temporal policy research.

Our hypothesis that urban areas produced the highest quality data was generally confirmed but there were examples of rural authorities that excelled at data collection and storage. We suggest this is largely due to the efforts made by individuals and teams within each authority. The most complete, highest quality data were received from three largest cities in Wales (Cardiff, Swansea and Newport) and resulted in a geocoding proportion of 66 %. Our proportion of geocoded urban outlets was lower than found in other research, which suggests 85 % geocoding may be achieved (Zandbergen 2008). However, their research focussed on residential addresses, which in our experience are more accurately recorded and more frequently contain core address components. We suspect the lack of structure in the commercial addresses we received contributed to our

lower geocoding success. Data received from the more sparsely populated rural unitary authorities were further investigated and for most fewer than half of the licensed premises provided could be geocoded. Given that there are statutory requirements to record these data and supply summary statistics to the Home Office we had expected that these licence data would be recorded more accurately. We recommend that communication is improved within and between unitary authorities to maximise the potential of these administrative data through initiatives such as the UK government funded Administrative Data Research Network (ADRN).

This study has used the most accurate geocoding standards to produce a complete spatial dataset of outlet locations by quarterly time-period at the highest resolution possible. Given that the quality of addresses varied systematically by rural and urban areas, the 97 % combined geocoding and manual match proportion achieved was the best solution for aggregation into small area geographies because it has reduced bias resulting from less accurately recorded rural outlets.

We suggest that the larger night-time economies in urban areas require urban unitary authorities to have accurate data on the location of premises from a planning and policing perspective. It is likely that this requirement has led to the manual adjustment of addresses by the urban authorities to match records when the licence application form is submitted for consideration. Moreover, there are likely to be more frequent checks on licensed premises in urban areas by licensing officers who update the records. Finally there is potentially greater business churn in urban areas leading to more frequent updating of records.

The use of a manual matching process to find the spatial location of licensed outlets is a relatively straightforward, if time consuming process. Limitations of the manual matching process include human error through misidentification of outlets from the address, and data errors remaining unresolved in Google Maps. However, the validation process has tested the introduction of errors with the reassuring result of 80 % located within 100 m of the geocoded location.

The spatial accuracy of the approximated locations (those which were not geocoded and failed to be manually matched precisely) is unknown and these outlets have been flagged in the dataset to enable removal for analyses, if so required. The full dataset will be used in our analyses for CHALICE because we are interested in changes in outlet density through time. Analyses concentrating on the spatial distribution of outlets within an urban area may choose to use only the geocoded and manually matched outlets. The geocoded, manually matched and approximated outlets have the greatest utility for subsequent data linkage at the household-level. These data will allow calculation of change in outlet density/accessibility through time at household- or LSOA-level to assess the impact on individual-level alcohol related health outcomes.

Longitudinal outlet data were required to assess quarterly changes through time and the impact on alcohol-related harm. The requirement to retrospectively collect data was challenging, with older premises more likely to have closed or changed name; changes were not always captured in official address registers such as AddressBase® Premium, despite the dataset structure to maintain historical property changes. The final dataset covered a six-year period and had consistent geocoding proportions throughout. The outlet data received from the unitary authorities seems not to have been subjected to any address verification, allowing local knowledge or colloquialisms into some alcohol outlet address data, probably by the landlord or licensee completing the form.

Occasionally the home or work address of the licensee or club secretary was recorded, rather than the outlet. These addresses are now ambiguous, leading to inaccuracies when trying to map an outlet, especially once it closed.

Standardised recording and maintenance of records between all unitary authorities could help to overcome these problems. A standardised address checking service may be incorporated into the address capture or data entry process. Manchester City Council, one of several comprising the Greater Manchester Unitary Authority in England, is achieving this by installing standardised address capture software within Council departmental software applications (John Bradley, Personal Communication). Councils should also be encouraged to maintain historical data, which are essential to retrospectively create and evaluate natural experiments.

Conclusions

The collation of retrospective alcohol outlet data was completed to enable the building of a longitudinal exposure dataset. There was considerable variation between the unitary authorities in the quality of address data and data related to the availability of alcohol, for example opening hours. The lack of address structure required us to devise a manual address matching process to capture the addresses that could not be geocoded. This extra procedure resulted in a geocoded alcohol outlet dataset suitable for creating an alcohol availability score at the household-level and any larger geography, including the LSOA-level. The alcohol outlet dataset will be used to contribute to the evidence-base of alcohol availability and alcohol-related harm.

The Licensing Act 2003 was intended to help Government monitor the impact of, among other things, the sale of alcohol in relation to alcohol-related crime. The potential of administrative licensing data as a resource for public health surveillance and associated data access issues have already been highlighted (Humphreys and Smith 2013). We faced similar problems trying to access Licence data. Through the process of geocoding a longitudinal dataset, we have identified further issues surrounding the variation in the number of data items stored, the accuracy of the address data, and historical preservation of the licence data. Moreover, many of the issues and problems we faced were raised more than 25 years ago in the Chorley report (Coppock 1987) and little progress has been made on “removing the barriers” to geographic information. Local government should be encouraged to use nationally standardised data fields, including addresses. The availability of accurate addresses, with dates, would contribute to research that aims to enhance our understanding of the impact of the built environment on health.

Acknowledgments We acknowledge the support from The Farr Institute CIPHER. The Farr Institute is supported by a 10-funder consortium: Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Council, the Medical Research Council, the National Institute of Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates), the Wellcome Trust (MRC Grant No: MR/K006525/1). This work was undertaken with the support of the Centre for the Development and Evaluation of Complex Interventions for Public Health Improvement (DECIPHer), a UKCRC Public Health Centre of Excellence. Joint funding (MR/KO232332/1) from the British Health Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, the Welsh Government and the Wellcome Trust, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged.

Compliance with Ethical Standards

Funding This project was funded by the National Institute for Health Research, Public Health Research (project number 09/3007/02). Please visit the PHR programme website for more information: <http://www.nets.nihr.ac.uk/projects/phr/09300702>. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Public Health Research, NIHR, NHS or the Department of Health.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Babor, T. F. (2008). Tackling alcohol misuse in the UK (editorial). *BMJ*, *336*, 455. doi:10.1136/bmj.39496.556435.80.
- British Medical Association (2008). *Alcohol misuse: tackling the UK epidemic*. London.
- Burgoine, T., & Harrison, F. (2013). Comparing the accuracy of two secondary food environment data sources in the UK across socio-economic and urban/rural divides. *International Journal of Health Geographics*, *12*, 2. doi:10.1186/1476-072X-12-2.
- Clarke, P., Ailshire, J., Melendez, R., Bader, M., & Morenoff, J. (2010). Using Google Earth to conduct a neighborhood audit: reliability of a virtual audit instrument. *Health & Place*, *16*, 1224–9. doi:10.1016/j.healthplace.2010.08.007.
- Coppock, J. T. (1987). Handling geographic information. report of the committee of enquiry chaired by Lord Chorley, Department of the Environment. (London: HMSO, 1987.) [Pp. 208.] Price £114-95. ISBN Oil 752015 2. *International Journal of Geographical Information Systems*, *1*, 191–192. doi:10.1080/02693798708927805.
- Cousens, S., Hargreaves, J., Bonell, C., Armstrong, B., Thomas, J., Kirkwood, B.R. et al. (2009). Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference. *Journal of Epidemiology: Community Health* jech.2008.082610. doi:10.1136/jech.2008.082610.
- Craig, P., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Ogilvie, D., Petticrew, M., Reeves, B., Sutton, M., & Thompson, S. (2012). Using natural experiments to evaluate population health interventions: new MRC guidance. *Journal of Epidemiology & Community Health*, *66*, 1182–1186. doi:10.1136/jech-2011-200375.
- Duncan, D. T., Castro, M. C., Blossom, J. C., Bennett, G. G., & Gortmaker, S. L. (2011). Evaluation of the positional difference between two common geocoding methods. *Geospatial Health*, *5*, 265–273.
- Egan, M., Kearns, A., Mason, P., Tannahill, C., Bond, L., Coyle, J., Beck, S., Crawford, F., Hanlon, P., Lawson, L., McLean, J., Petticrew, M., Sautkina, E., Thomson, H., Walsh, D., & GoWell Team. (2010). Protocol for a mixed methods study investigating the impact of investment in housing, regeneration and neighbourhood renewal on the health and wellbeing of residents: the GoWell programme. *BMC Medical Research Methodology*, *10*, 41. doi:10.1186/1471-2288-10-41.
- Ellaway, A., Macdonald, L., Forsyth, A., & Macintyre, S. (2010). The socio-spatial distribution of alcohol outlets in Glasgow city. *Health & Place*, *16*, 167–172.
- Flatley, J., Kershaw, C., Smith, K., Chaplin, R., & Moon, D. (2010). *Crime in England and Wales 2009/10*. London: Home Office.
- Fone, D., Dunstan, F., White, J., Webster, C., Rodgers, S., Lee, S., Shiode, N., Orford, S., Weightman, A., Brennan, I., Sivarajasingam, V., Morgan, J., Fry, R., & Lyons, R. (2012). Change in alcohol outlet density and alcohol-related harm to population health (CHALICE). *BMC Public Health*, *12*, 428. doi:10.1186/1471-2458-12-428.
- Fone, D. L., Dunstan, F., White, J., Kelly, M., Farewell, D., John, G., Lyons, R. A., & Lloyd, K. (2013). Cohort profile: the Caerphilly Health and Social Needs Electronic Cohort Study (E-CATALyST). *International Journal of Epidemiology*, *42*, 1620–1628. doi:10.1093/ije/dys175.
- ONS Geography (2004). Rural/Urban Definition (England and Wales).

- Gruenewald, P. J. (2011). Regulating availability: how access to alcohol affects drinking and problems in youth and adults. *Alcohol Research & Health*, *34*, 248–256.
- Hadfield, P., Lister, S., & Traynor, P. (2010). “This town”s a different town today’. *Criminology & Criminal Justice*, *9*, 465–485. doi:10.1177/1748895809343409.
- Halonen, J. I., Kivimäki, M., Virtanen, M., Pentti, J., Subramanian, S. V., Kawachi, I., & Vahtera, J. (2013). Living in proximity of a bar and risky alcohol behaviours: a longitudinal study. *Addiction*, *108*, 320–328. doi:10.1111/j.1360-0443.2012.04053.x.
- Humphreys, D. K., & Smith, D. M. (2013). Short report. alcohol licensing data: why is it an underused resource in public health? *Health & Place*, *24*, 110–114. doi:10.1016/j.healthplace.2013.07.006.
- Klatsky, A. L., & Gunderson, E. (2008). Alcohol and hypertension: a review. *Journal of the American Society of Hypertension*, *2*, 307–317.
- Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? evaluating the accuracy of geocoding in public health research. *American Journal of Public Health*, *91*, 1114–6. doi:10.2105/AJPH.91.7.1114.
- Leon, D. A., & McCambridge, J. (2006). Liver cirrhosis mortality rates in Britain from 1950 to 2002: an analysis of routine data. *Lancet*, *367*, 52–56. doi:10.1016/S0140-6736(06)67924-5.
- Ogilvie, D., Mitchell, R., Mutrie, N., Pettecrew, M., & Platt, S. (2006). Evaluating health effects of transport interventions: methodologic case study. *American Journal of Preventive Medicine*, *31*, 118–126. doi:10.1016/j.amepre.2006.03.030.
- Ordnance Survey (2014). AddressBase Premium | Business and government | Ordnance Survey [WWW Document]. URL <http://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-premium.html> (accessed 3.27.14).
- Parkin, D. M. (2011). Cancers attributable to consumption of alcohol in the UK in 2010. *British Journal of Cancer*, *105*(Suppl 2), S14–18. doi:10.1038/bjc.2011.476.
- Pateman, T. (2011). Rural and urban areas: comparing lives using rural/urban classifications. *Regular Trends*, *43*, 11–86. doi:10.1057/rt.2011.2.
- Pettecrew, M., Cummins, S., Ferrell, C., Findlay, A., Higgins, C., Hoy, C., Kearns, A., & Sparks, L. (2005). Natural experiments: an underused tool for public health? *Public Health*, *119*, 751–757.
- Pettecrew, M., Kearns, A., Mason, P., & Hoy, C. (2009). The SHARP study: a quantitative and qualitative evaluation of the short-term outcomes of housing and neighbourhood renewal. *BMC Public Health*, *9*, 415. doi:10.1186/1471-2458-9-415.
- 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 and Alcoholism*, *44*, 500–516.
- Ramstedt, M. (2001). Alcohol and suicide in 14 European countries. *Addiction*, *96*, 59–75.
- Ratcliffe, J. H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*. doi:10.1080/13658810310001596076.
- Reynolds, K., Lewis, L. B., Nolen, J. D. L., Kinney, G. L., Sathya, B., & He, J. (2003). Alcohol consumption and risk of stroke: a meta-analysis. *JAMA*, *289*, 579–588.
- Richardson, E., Shortt, N., & Pearce, J. (2014) Alcohol outlets and health in Scotland. Alcohol Focus Scotland, Centre for Research on Environment, Society and Health, CRESH.
- Rodgers, S. E., Heaven, M., Lacey, A., Poortinga, W., Dunstan, F. D., Jones, K. H., Palmer, S. R., Phillips, C. J., Smith, R., John, A., Davies, G. A., & Lyons, R. A. (2012). Cohort profile: the housing regeneration and health study. *International Journal of Epidemiology*, *43*, 52–60. doi:10.1093/ije/dys200.
- Rossen, L. M., Pollack, K. M., & Curriero, F. C. (2012). Verification of retail food outlet location data from a local health department using ground-truthing and remote-sensing technology: assessing differences by neighborhood characteristics. *Health & Place*, *18*, 956–62. doi:10.1016/j.healthplace.2012.06.012.
- Royal Mail (2010). Royal Mail Programmers Guide Edition 7, version 2.3 [WWW Document]. URL http://www.poweredbypaf.com/wp-content/themes/amu/paf_downloads/programmers_guide.pdf (accessed 3.27.14).
- Rundle, A. G., Bader, M. D. M., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to audit neighborhood environments. *American Journal of Preventive Medicine*, *40*, 94–100. doi:10.1016/j.amepre.2010.09.034.
- Local Government Data Unit - Wales (2005). Welsh index of multiple deprivation 2005. Summary Report. [WWW Document]. LGDU.
- Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., Mix, W., Colt, J. S., & Hartege, P. (2005). Positional accuracy of two methods of geocoding. *Epidemiology Cambridge Massachusetts*, *16*, 542–7.

- White, J., Greene, G., Dunstan, F., Rodgers, S., Lyons, R. A., Humphreys, I., John, A., Webster, C., Palmer, S., Elliott, E., Phillips, C. J., & Fone, D. (2014). The Communities First (ComFi) study: protocol for a prospective controlled quasi-experimental study to evaluate the impact of area-wide regeneration on mental health and social cohesion in deprived communities. *BMJ Open*, *4*, e006530. doi:[10.1136/bmjopen-2014-006530](https://doi.org/10.1136/bmjopen-2014-006530).
- Xu, Y., Yu, Q., Scribner, R., Theall, K., Scribner, S., & Simonsen, N. (2012). Multilevel spatiotemporal change-point models for evaluating the effect of an alcohol outlet control policy on changes in neighborhood assaultive violence rates. *Spatio Spatio-Temporal Epidemiology*. doi:[10.1016/j.sste.2012.04.005](https://doi.org/10.1016/j.sste.2012.04.005).
- Young, R., Macdonald, L., & Ellaway, A. (2013). Associations between proximity and density of local alcohol outlets and alcohol use among Scottish adolescents. *Health & Place*, *19*, 124–130. doi:[10.1016/j.healthplace.2012.10.004](https://doi.org/10.1016/j.healthplace.2012.10.004).
- Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health*, *7*, 37. doi:[10.1186/1471-2458-7-37](https://doi.org/10.1186/1471-2458-7-37).
- Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, *32*, 214–232. doi:[10.1016/j.compenvurbsys.2007.11.006](https://doi.org/10.1016/j.compenvurbsys.2007.11.006).