Article:
Andresen, MA, Malleson, N orcid.org/0000-0002-6977-0615, Steenbeek, W et al. (2 more authors) (2020) Minimum geocoding match rates: an international study of the impact of data and areal unit sizes. International Journal of Geographical Information Science, 34 (7). pp. 1306-1322. ISSN 1365-8816

https://doi.org/10.1080/13658816.2020.1725015

Minimum geocoding match rates: an international study of the impact of data and areal unit sizes

Abstract

The analysis of geographically referenced data, specifically point data, is predicated on the accurate geocoding of those data. Geocoding refers to the process in which geographically referenced data (addresses, for example) are placed on a map. This process may lead to issues with positional accuracy or the inability to geocode an address. In this paper, we conduct an international investigation into the impact of the (in)ability to geocode an address on the resulting spatial pattern. We use a variety of point data sets of crime events (varying numbers of events and types of crime), a variety of areal units of analysis (varying the number and size of areal units), from a variety of countries (varying underlying administrative systems), and a locally-based spatial point pattern test to find the levels of geocoding match rates to maintain the spatial patterns of the original data when addresses are missing at random. We find that the level of geocoding success depends on the number of points and the number of areal units under analysis, but generally show that the necessary levels of geocoding success are lower than found in previous research. This finding is consistent across different national contexts.

Keywords: geocoding; match rate; accuracy; modifiable areal units; spatial point pattern test

Introduction

Geocoding spatially referenced data, regardless of the context, is often the first step towards a spatial analysis. Though many sources of spatial data are now provided to researchers with coordinates such that the data can be mapped immediately—e.g., spatially referenced crime data are almost ubiquitous—geocoding algorithms are always operating in these contexts either in the background or explicitly when researchers have to geocode spatially-referenced data themselves. However, it is important to note that research has shown that geocoding algorithms are not only inaccurate at times—such that the geocoded events are not placed in the correct spatial location—but are also at risk of not being able to locate some street addresses or street intersections for events in the first place (Ratcliffe 2001; Cayo and Talbot 2003; Zandbergen 2008).

The current analysis investigates the impact of not being able to geocode a subset of a spatially-referenced data set, i.e. *match rates*. Data not being geocoded may result from any number, or combination, of situations that include spelling mistakes, incorrect references or abbreviations in street types, impossible addresses, and missing information (Ratcliffe 2004). Because of the importance of accuracy in spatial data (Bailey and Gatrell 1995; O'Sullivan and Unwin 2010), data that are not geocoded may have significant implications for subsequent research. Specifically, missing data, even missing at random, may lead to bias in the spatial pattern of the events being analyzed. As administrative systems vary internationally, it is also possible that some national jurisdictions produce data that are often (in)sufficient for accurate spatial analysis at small areas.

But how much data can be missing before the spatial pattern of the mapped data becomes insufficiently similar to the complete data? Research in this area is limited; in fact,

we are only aware of one research paper that has investigated this issue, which only focused on New South Wales, Australia (Ratcliffe 2004). In this paper, we contribute to this literature by conducting an international investigation using a variety of data sets with different counts of events, multiple and different-sized areal units, from multiple cities and countries, and a locally-based spatial point pattern test. We are able to show, with consistency across cities and countries, that when data are missing at random, the minimum acceptable match rate is context dependent, specifically with regard to the size and number of areal units under analysis. Importantly, the results suggest that the necessary levels of geocoding success are lower than found in previous research.

Related research

Overall, there are a number of potential problems that may arise in geocoding:

- long streets may be arbitrarily broken into segments that are not based on intersections;
- events can be placed on the street segment using an interpolation process that may place the event in the wrong place on the street segment;
- a geocoding match may be made on an areal unit and subsequently misplaced on the wrong street segment;
- there is variation in street segment length that may skew the analysis; and
- the geocoding process may fail to find a location for an event at all (match rate) (Chainey and Ratcliffe 2005).

There are also privacy concerns with regard to reverse geocoding, particularly for sensitive data related to crime and public health (Kounadi *et al*. 2013). There is a significant body of

research that has investigated these issues, some of which is reviewed below, as well as other research that highlights the importance of data quality (Bichler and Balchak 2007), alternative geocoding services and techniques (Bell *et al*. 2012Murrey *et al*. 2011; Whitsell *et al*., 2006), and the advantages of various online geocoding services in a variety of different research contexts (Roongpiboonsopit and Karimi 2010a, 2010b; Davis and Alencar 2011; Karimi *et al*. 2011; Mazeika and Summerton 2017).

The component of the geocoding literature most pertinent to the current topic, however, aside from research on match rates, is the accuracy (positional error) of geocoded data: relative to the actual address, where do geocoding procedures[1] place the point on the map? This research area is of particular importance in research that considers the geography of health with issues for health access, risk, and outcomes—see Goldberg and Jacquez (2012) for a special issue on geocoding and positional accuracy that extend beyond the current scope of this paper.

Research has shown that even slight levels of positional error can have notable impacts on subsequent analyses (Malizia 2013), such that accuracy in geocoding may matter for both data that are and are not placed on a map, particularly in the case of spatial cluster detection (Zimmerman et al., 2008). This is particularly relevant given the recent importance attributed to *micro-places* in crime analysis (Weisburd 2015; Steenbeek and Weisburd 2016), i.e. precise locations matter. Moreover, it was not that long ago when spatially-referenced data were organized and analyzed without the aid of computers, particularly in the case of crime analysis (Chainey and Ratcliffe 2005).

[1] See Goldberg (2011) for a discussion of the methods for geocoding address data in geographic information systems.

Investigating the impact of geocoding addresses onto a street segment using linear interpolation (150 Main Street is placed in the middle of the 100-block of Main Street, for example), Ratcliffe (2001) found that an address is placed in an incorrect census tract 5 – 7.5 percent of the time.[2] Given the result from Malizia (2013), even this low degree of positional inaccuracy may be problematic for subsequent analyses. More problematic, however, is when analyses are undertaken at units of analysis cartographically larger than the unit of analysis used in geocoding: Ratcliffe (2001) found that more than 50 percent of points were placed within the incorrect land parcel. Though such inaccuracies may be considered discouraging for those analyzing spatial point patterns, this issue can be avoided by not undertaking inference at a spatial scale "lower" than data quality can justify, similar to the primary strategy for avoiding the ecological fallacy (Openshaw 1984).

Cayo and Talbot (2003), also investigating positional inaccuracies, found that geocoding to the parcel rather than to the street segment resulted in greater accuracy, an intuitively sensible result—see Zandbergen (2008) for an exception. Moreover, they found that positional accuracy was better in urban areas relative to suburban areas and suburban areas relative to rural areas. This should come as no surprise simply because of the decreasing degree of street density, and the corresponding increases in average error, moving away from urban centres.

More recently, in an analysis using a wide variety of crime types, with the number of data points ranging from approximately 600 to 100,000 and different road network files,

[2] Linear interpolation is most common in North America, given that European streets tend not to be regular enough for interpolation to be useful.

Hart and Zandbergen (2013) found that different crime types and road networks had an impact on geocoding match rates and positional accuracy. As such, data quality (addresses being geocoded and the spatial reference data), is an important consideration, but they also found that geocoding to larger areal units (street segments versus actual addresses or parcels) led to better match rates and higher levels of positional accuracy—Shah et al. (2014) found comparable results. Moreover, Edwards et al. (2014) found similar results with better match rates using larger areal units, but that these match rates are also better in urban and higher incomes areas, analogous to those found by Cayo and Talbot (2003) with regard to urban areas.

Most pertinent to the current investigation is Ratcliffe (2004) who identified the minimum acceptable match rate in geocoding that must be achieved when data are missing at random for the spatial pattern of the mapped data to be the same as, or similar enough to, the complete data. In this paper, Ratcliffe (2004) uses a Monte Carlo approach to identify statistically significant differences in the spatial patterns of geocoded (crime) data. First, 100 percent of the geocoded data are assigned to areal units, and those areal units are ranked according to the highest and lowest counts. Second, 1 percent of the data are removed at random, and then the areal units are again ranked according to the highest and lowest counts. Third, using the Mann-Whitney U test (a nonparametric rank test), the differences in ranks are tested statistically. Fourth, if the difference is statistically significant stop, but if not repeat the second and third steps until a statistically significant difference is found. Fifth, repeat the above 250 times in order to be sure that geocoding match rates are not identified based on one (or a few) aberrant samples; the mean of these

simulations in each case, plus 2 standard deviations, was used to identify the minimum acceptable match rate.

Ratcliffe (2004) undertook this simulation using a variety of crime types (all reported crime, vehicle crime, malicious damage, and burglary ranging in counts from 783 to 1362) and areal units (census blocks ranging in count from 144 to 261) in five different areas of New South Wales, Australia. In the case with the greatest number of points, Ratcliffe (2004) identified 78 percent as minimum acceptable match rate, but either 84 or 85 percent in the other 4 cases. In order to err on the side of caution, 85 percent was identified as the overall minimum acceptable match rate that has been cited in many studies that have used geocoding procedures since its publication—over 300 citations at the time of writing—most often in criminology but also geography, (public) health, and epidemiology. [3]

Ratcliffe (2004) identifies a number of limitations in his analysis. First and foremost, as repeatedly stated in the article, this is a *first* estimate. As such, more research is needed in this area. Though different data sets in different contexts are used, this is only one study in one country. There are two important limitations in these analyses: (1) the variation in the number of data points and areal units; and (2) the test used for statistical testing. First, though five data sets in different contexts are used, the range for event data is only 600 events: 783 to 1362. In standard statistical sampling, the sample size required for a representative sample effectively goes asymptotic after a point. This may be true for geocoding events as well, such that after a certain point, if data are truly missing at random, the minimum acceptable match rate decreases as the total event population increases. Also,

[3] https://scholar.google.com/scholar?oi=bibs&hl=en&cites=8644180519281685630

the number of areal units only range from 144 to 261. Though this may be representative of census tracts, or similarly sized areal units, Ratcliffe (2004) did not investigate smaller areal units such as census block groups. Moreover, the crime and place literature advocates street segments, using tens of thousands of areal units in their analyses (see Andresen *et al*. 2017a, 2017b, Bernasco and Steenbeek 2017; Braga *et al*. 2010, 2011; Vandeviver and Steenbeek 2019; Weisburd 2015; Weisburd *et al*. 2004, 2012; Wheeler *et al*. 2016); surely, the number of areal units in the analysis impacts the need for a greater or lesser minimum acceptable match rate. And second, though identified by Ratcliffe (2004) as having other limitations, the Mann-Whitney U test is a global statistic. As such, one spatial pattern could be very clustered and another could be very close to uniform, but have identical rankings, not identifying any statistically significant differences.

In order to address these limitations, in the analyses below we consider a much wider range of event counts, ranging from less than one thousand to over 10,000. We also consider a wider range of areal units of analysis, ranging from less than 100 up to almost 50,000, therefore being representative for neighborhood level analyses as well as those considering the micro-place. We use a locally-based spatial point pattern test that identifies statistically significant change for each areal unit, not the overall ranking. And finally, though we use five locations, similar to Ratcliffe (2004), these five locations are in different (Western) countries and represent very different underlying physical environments (e.g., North-American-style regular grids compared to more organically evolved European-style street networks).

Materials and methods

Study sites, data, and areal units

The cities (countries) we investigate geocoding match rates are Vancouver (Canada), Leeds (England), The Hague (Netherlands), Brisbane (Australia), and Antwerp (Belgium). As shown in Table 1, a variety of property and violent crime types are used for events to investigate geocoding match rates. Having a variety of event types is not the goal, however, but to have a range of sizes for the data sets under investigation. We selected, where possible, event classifications that had approximately 1,000, 2,500, 5,000, and 10,000 events. We always selected complete data sets within each event type, rather than (randomly) selecting specific sample sizes. With regard to areal units[4], for each city we selected three areal units of analysis, often defined by national censuses or local administrative boundaries. Local and national conventions define the size of these various areal units, but land area for each city is also related to their counts.

In all cases, we have a variety for the number of areal units to investigate their impact on geocoding match rates. And with the exception of Leeds, we are able to have a range of event counts to investigate the potential impact of minimal acceptable match rates decreasing as the number of events increases. Combined, we investigate 54 event-unit combinations.

<Insert Table 1 About Here>

Table 1. Counts of events (crime) and areal units of analysis

[4] In order to undertake the spatial point pattern test, outlined below, street segments include non-overlapping buffers (7 metres) such that these line features are areal units for the purposes of the test.

Spatial point pattern test and the Monte Carlo simulation

In order to address our research question regarding the minimum acceptable match rate, we need to use a spatial statistical test that can identify change at the local level. For this purpose, we use the spatial point pattern test developed by Andresen (2009, 2016), and extended by Steenbeek *et al*. (2018) and Wheeler *et al*. (2018). This spatial point pattern test identifies spatial stability and/or differences in two (or more) spatial point patterns. This test is undertaken considering the percentages of event types (crime) in each areal unit of analysis and, therefore, can identify differences between different level of geocoding. A review of the details and applications of the spatial point pattern test is available in Wheeler *et al.* (2018), with the most recent applications being in the contexts of comparing spatial patterns in forensic DNA data and police recorded crime data (De Moor *et al*. 2018), the changing spatial patterns of crime with regard to the crime drop (Hodgkinson and Andresen 2019; Hodgkinson *et al*. 2016; Vandeviver and Steenbeek 2019), the spatial dimension of police proactivity (Wu and Lum 2017), and the appropriate spatial scale for aggregate crime analysis (Malleson *et al*. 2019).

The spatial point pattern test identifies differences in the spatial patterns of two, or more, point data sets considering an underlying areal unit of analysis; as such, this is an areal-based spatial point pattern test The output of this test is a global index of similarity, *S*, that ranges between 0 (no similarity) and 1 (perfect similarity), calculated as follows:

$S = \frac{\sum_{i=1}^{n} s_i}{n}$	(1)

where s_i is equal 1 if the pattern of two datasets are similar within an individual spatial unit of analysis and 0 otherwise, and *n* is the number of areas. As such, the *S*-Index measures the

percentage of areas (street segments or census tracts, for example) that share a similar spatial pattern. An *S*-Index value of 0.80 or greater is often used to identify when two spatial point patterns are similar (Andresen 2009), however, we err on the conservative side and use 0.90, but this has little qualitative impact on the results presented below. In addition to the *S*-Index, the results of this test may be mapped, allowing for local level results to be shown and, subsequently, analyzed for their patterns. There are a number of versions of this test available with the most relevant being a full bootstrap, "partial" bootstrap, and a proportion difference test.

In the context of two spatial point patterns, the full bootstrap version treats both data sets as random realizations of known spatial patterns, undertaking a full bootstrap with replacement on both data sets. The partial bootstrap version identifies one of the spatial point patterns as a base and the other as a test; the base is considered a set of known values (percentage of points geocoded to an areal unit, for example) and the test data set is treated as a random realization and has a full bootstrap with replacement. Lastly, the proportion difference test identifies statistically significant change using chi-square tests at the local level rather than a nonparametric confidence interval. We use the partial bootstrap version of the test in the Monte Carlo simulation below, because we treat the percentages of events for each areal unit in the original event data sets as known (the base data sets) and perform a bootstrap on the subsequently sampled data sets as the test data sets. All versions of the test are available as an R library (Steenbeek *et al*. 2018)—see

Andresen (2009, 2016), Wheeler *et al*. (2018), and Steenbeek *et al*. (2018) for more details regarding the test options.[5]

The basic context of the spatial point pattern test is as follows:

1. Identify one data set as the base and calculate the percentage of events within each areal unit;
2. Randomly sample with replacement the test data set and calculate the percentage of events within each areal unit;
3. Undertake a Monte Carlo simulation by repeating step 2 a number of times (we use 200) to generate a confidence interval of values for each areal unit;
4. Compare the percentage from each areal unit in the base data set to its corresponding confidence interval (we use 95 percent), such that if the base value is within the confidence interval it is considered similar;
5. Calculate the *S*-Index as the percentage of areal units that are considered similar, with 0.80 or greater indicating similarity.

We adapt this general procedure by undertaking these steps with successively smaller sample of test data, one percent at a time, similar to Ratcliffe (2004). As such, we randomly select 99 percent of the base data for the test data and then complete steps 1 through 5, then 98 percent, 97, 96, 95, and so on until we reach 1 percent. This simulates randomly missing data, removing 1 percent at a time. We conduct this procedure 10 times for each percentage (99 through to 1 percent), for a total of 2,000 simulations for each combination

[5] We will publish the full code for replication purposes in a GitHub repository with a link in this footnote, currently suppressed for the review process.

(using more than ten replications does not change the substantive results reported here). The output from these simulations is 54 sets of *S*-Index – geocoding match rate combinations.

Results

The primary results from the simulations are shown for each city in Figures 1 through 5; each figure has the three difference scales for each crime type in its respective city and a horizontal line to represent an S-Index value of 0.90. Though there are differences across cities and crime types, there is one general set of results. First, the relationship between the *S*-Index and geocoding match rates for the smallest areal units of analysis (street segments or mesh blocks) is approximately linear with $S >= 0.90$ emerging around a 85 percent match rate. Second, the medium- (or middle-) sized areal units tend to reach $S >= 0.90$ when geocoding match rates are 50 to 70 percent. Third, the large areal units tend to reach $S >= 0.90$ with geocoding hit ranges ranging from 10 to 30 percent. And fourth, in all cases the larger areal units have greater *S*-Index values for a given geocoding match rate, followed by the medium-sized areal units, and the smallest areal units.

<Insert Figures 1 to 5 About Here>

Figure 1. S-Index – geocoding match rate, Vancouver

Figure 2. S-Index – geocoding match rate, Leeds

Figure 3. S-Index – geocoding match rate, The Hague

Figure 4. S-Index – geocoding match rate, Brisbane

Figure 5. S-Index – geocoding match rate, Antwerp

Table 2 shows the geocoding match rate (to the closest 5 percent level, rounded up) to achieve an *S*-Index value >= 0.90. There is remarkable consistency for the small areal units of analysis across all crime types and cities: a geocoding match rate of 85 to 90 percent is necessary to achieve this threshold *S*-Index value. The medium-sized areas have notably smaller geocoding match rates necessary to achieve an *S*-Index value >= 0.90. Moreover, a general pattern emerges for each city that is not easily discernible from the figures: as the number of events increase for different crime types within each city, the necessary geocoding match rate decreases—this is somewhat evident, but not as immediately obvious for the smallest areal units. Therefore, holding the number of areal units constant, as the number of events increases, the geocoding match rate necessary to maintain the relative spatial pattern decreases, as hypothesized above. In other words, there are diminishing marginal returns to increases in the number of events when geocoding given an areal dataset. This is also present for the larger areal units of analysis.

There is a general pattern of requiring lower geocoding match rates when there are increases in the number of events and/or decreases in the number of areal units. However, it is difficult to discern a clear relationship in this context because of the varying counts of events and the different numbers (particularly relative to city size) of areal units under analysis. In order to examine this complexity, Figure 6 plots the geocoding match rate necessary to achieve an *S*-Index >= 0.90 relative to the ratios of events-areas (Figure 6a) and the natural logarithm of areas-events (Figure 6b)—Loess regressions (a nonparametric smoothing technique) are shown in both figures to isolate the trends.

<Insert Table 2 About Here>

Table 2. Geocoding match rate, *S*-Index = 0.90, crime type and area type

Figure 6a shows a quadratic trend with the pattern: as the number of events increases relative to the number of areal units, the necessary geocoding match rate to achieve an *S*-Index >= 0.90 decreases, albeit nonlinearly. However, with this ratio ranging from close to zero to 250, it is somewhat difficult to interpret. Figure 6b shows the natural logarithm of the ratio of the number of areal units to the number of events. As such, a value of zero indicates that the number of areal units is equal to the number of events and as this value increases, this indicates an increase in the number of units, with the upper extreme being street segments and mesh blocks. Figure 6b also includes a vertical line when the ratio is equal to zero and a horizontal line when the geocoding match rate is 80 percent. This all shows that when the ratio of areal units is equal to or greater than the number of events a geocoding match rate of 80 percent or greater is necessary to achieve an *S*-Index >= 0.90. Conversely, the necessary geocoding match rate decreases steadily as the number of events is greater than the number of areal units. These relationships have remarkably good fits with their respective R^2 values being 0.87 (modelled with a quadratic term) and 0.94, respectively.

<Insert Figure 6 About Here>

Figure 6. Geocoding match rate and ratio of events to areas

Discussion

We have shown that the previously suggested 85 percent acceptable minimum geocoding match rate (Ratcliffe, 2004) holds in a particular situation, but cannot simply be generalized to any situation or context. An 85 percent geocoding match rate is necessary to maintain spatial patterns when the analysis used street segments, or their equivalent size,

as the spatial unit of analysis. This result is particularly pertinent, then, to those undertaking research within the crime and place literature who use micro-places for their units of analysis (Weisburd 2015). Moreover, whenever there are more areas than events, a minimum geocoding match rate of 80 percent is necessary. As such, in order to err on the side of caution and maintain a conservative approach, Ratcliffe's (2004) 85 percent minimum acceptable geocoding match rate could always be applied when there are more areas than events.

However, when the number of events begins to be greater than the number of areas, the minimum geocoding match rate does not need to be as conservative. In fact, once event-area ratio reaches 10 (1000 events and 100 areal units, for example) a geocoding match rate of 50 percent is sufficient to achieve an *S*-Index >= 0.90, and this scenario is not an uncommon situation even for relatively rare crime types in a moderately sized city. A natural question to ask at this point is why Ratcliffe (2004) found that an 85 percent geocoding match rate was necessary when the event-area ratios ranged from 4 to 9 in their research? Though we cannot say with confidence without analyzing their data, we are confident with our results because we analyzed many contexts with consistency across crime types, the number of spatial units of analysis, and different cities/countries.

It is important to note here that when using the larger areal units, many of our analyses showed that a 20 to 30 percent geocoding match rate (or less) was sufficient to achieve an *S*-Index >= 0.90. Though this may seem rather low and have the reader concerned about any analyses that use spatially-referenced data with such a low geocoding match rate, we must recall that the current analysis is concerned with data that are missing at random. Eventually, randomly missing data do lead to changes in the spatial patterns

(see Ratcliffe 2004), but surely 20 percent is too low a threshold even when considering "standard" areal units of analysis (census tracts) and a sufficiently large event type data set?

We would raise serious concerns regarding any data set that had such a low geocoding match rate, questioning if data are truly missing at random, but suppose that they are. Many national level censuses have two data gathering processes: a "full census" in which every household fills out a census form about every individual in the household for a limited number of questions, and a "partial census" in which approximately 20 percent of households fill our a much more in depth form to obtain socio-economic and socio-demographic information. These partial censuses have a sample size of 20 percent (this may vary slightly from country to country) because when they are random they have the expected values from the full population with low enough standard errors to make meaningful inferences for research and public policy. As such, if there are a sufficient number of events and not that many areas, it should come as no surprise that a 20 percent geocoding match rate (or even lower) is sufficient to achieve an *S*-Index >= 0.90 if the data are truly missing at random. The difficulty, of course, is being able to properly assess such a situation.

Regardless, the important result here is that the 85 percent minimum acceptable match rate, though relevant in some contexts, particularly in the crime and place literature, is not a monolithic result. Once there are more events than areas, geocoding match rates less than the 85 percent threshold are acceptable for the maintenance of spatial patterns if data can be identified as missing at random. In fact, this requirement of randomness for missing data was stated by Ratcliffe (2004) for his 85 percent geocoding match rate as well.

Though claims of acceptable geocoding match rates must always be made with this caution in mind, one should not consider lower geocoding match rate thresholds if they are unable to properly inspect their missing data. If such an inspection is not possible (XY coordinates are just missing with no addresses to check) or not feasible (too many events to manually inspect the data) then even the 85 percent minimum acceptable geocoding match rate may not be enough because research on geocoding has not investigated this phenomenon.

As with all research, ours is not without limitations. First, though we have 54 event-area combinations for our simulations that leads to a strong relationship between geocoding match rates and the area-event ratio, many more possibilities could be investigated. Specifically, this simulation could be extended to increasingly expand, or shrink a regular grid over the various study areas to see if more area-event ratios can shed more light on this relationship. Second, we only consider criminal event data. Many other disciplines such as (public) health, economics, sociology, and political science use event level data and it is possible that the relationships found here will not generalise to those event types. And third, we only consider data that are missing at random. Though this may be a common reason for the inability to geocode data, as discussed above (misspelled streets, impossible addresses, improper street types, and so on), such data entry errors may be systematic in some cases. As such, the geocoding match rates referred to here must be used with caution and, as stated by Ratcliffe (2004) in their research, only when missing data appear to be missing at random.

In addition to addressing the limitations above, there are a number of directions for future research, with the most obvious being related to the nature of non-geocoded data. For example, in some policing jurisdictions if an exact address for a criminal event is not

known but the neighborhood is, the "address" for the criminal event may be assigned as the centroid of the neighborhood or a random point within that neighborhood. Additionally, if a criminal event has an unknown location, aside from the city within which it occurred, that criminal event may be assigned to the police station or substation where it was reported. And third, there may be systematic errors in criminal event location reporting such that the same places are consistently not geocoded. The first two directions for future research are tractable to investigate with a point level data set and areal boundary data, but the third is more difficult. Specifically, a complete set of events with addresses would have to be obtained with a subsequent data set of geocoded events. The non-geocoded events would then have to be identified (relatively easy) and then it is necessary to search for a pattern that led to unsuccessful geocoding output (more difficult). The difficulty is that many spatially-referenced data sets are provided without specific addresses but with variables for X and Y coordinates. If an event is not geocoded it simply has missing values for the XY coordinates without the address, often for confidentiality concerns.

Regardless of this difficulty, there is still a lot of research necessary to investigate geocoding match rates. As spatially-referenced data sets are increasingly become available, spatial analyses of such data are more and more commonplace. Given the added dimensions, literally, of spatial data, acknowledging its quality is critical if we are to make theoretical advancements in understanding social phenomena and the development of (public) policy to improve social ills.

Data and codes availability statement

The codes that support the findings of this study are available in “figshare.com” with the identifier(s) at the private link (https://figshare.com/s/fa33812e909b0f71e049).

All crime data cannot be made publicly available due to confidentiality agreements with the respective policing agencies.

Acknowledgments

The authors gratefully acknowledge use of the services and facilities of the Griffith Criminology Institute's Social Analytics Lab at Griffith University. Police data used in this research have been extracted from the Griffith Social Analytics Laboratory and has not been centrally verified by the Queensland Police Service. Responsibility for any errors of omission or commission remains with the author(s). The Queensland Police Service expressly disclaims any liability for any damage resulting from the use of the material contained in this publication and will not be responsible for any loss, howsoever arising, from use of or reliance on this material.

Declaration of interest statement

No potential conflict of interest was reported by the authors.

References

Andresen, M. A., 2009. Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. *Applied Geography*, 29(3), 333–345.

Andresen, M.A., 2016. An area-based nonparametric spatial point pattern test: The test, its applications, and the future. *Methodological Innovations*, 9, Article 12.

Andresen, M. A., Curman, A. S. N., and Linning, S. J., 2017a. The trajectories of crime at places: Understanding the patterns of disaggregated crime types. *Journal of Quantitative Criminology*, 33(3), 427 - 449.

Andresen, M. A., Linning, S. J., and Malleson, N., 2017b. Crime at places and spatial concentrations: Exploring the spatial stability of property crime in Vancouver BC, 2003-2013. *Journal of Quantitative Criminology*, 33(2), 255 - 275.

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. Harlow, UK: Prentice Hall.

Bell, S., Wilson, K., Shah, T. I., Gersher, S., & Elliott, T. (2012). Investigating impacts of positional error on potential health care accessibility. *Spatial and Spatio-temporal Epidemiology*, 3(1), 17-29.

Bernasco, W., and Steenbeek, W., 2017. More places than crimes: Implications for evaluating the law of crime concentration at place. *Journal of Quantitative Criminology*, 33(3), 451 – 467.

Braga, A., Hureau, D. M., and Papachristos, A. V., 2010. The concentration and stability of gun violence at micro places in Boston, 1980–2008. *Journal of Quantitative Criminology*, 26(1), 33 – 53.

Braga, A., Hureau, D. M., and Papachristos, A. V., 2011. The relevance of micro places to

citywide robbery trends: A longitudinal analysis of robbery incidents at street corners and block faces in Boston. *Journal of Research in Crime and Delinquency*, 48(1), 7 – 32.

Cayo, N. R., and Talbot, T. O., 2003. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics*, 2, 1 – 12.

Chainey, S., and Ratcliffe, J.H., 2005. *GIS and crime mapping*. Chichester, UK: John Wiley & Sons.

Davis Jr., C. A., and de Alencar, R. O., 2011. Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city. *Transactions in GIS*, 15(6), 851 – 868.

De Moor, S., Vandeviver, C., and Vander Beken, T., 2018. Are DNA data a valid source to study the spatial behavior of unknown offenders? *Science and Justice*, 58(5), 315-322.

Edwards, S. E., Strauss, B., and Miranda, M. L., 2014. geocoding large population-level administrative datasets at highly resolved spatial scales. *Transactions in GIS*, 18(4), 586–603.

Goldberg, D. W., 2011. Improving geocoding match rates with spatially-varying block metrics. *Transactions in GIS*, 15(6), 829 – 850.

Goldberg, D. W., & Jacquez, G. M. [eds.] (2012). Special issue on geocoding in the health sciences. *Spatial and Spatio-temporal Epidemiology*, 3(1), 1-92.

Hart, T. C., and Zandbergen, P. A., 2013. Reference data and geocoding quality. *Policing: An International Journal of Police Strategies and Management*, 36(2), 263 – 294.

Hodgkinson, T., and Andresen, M.A. (2019). Changing spatial patterns of residential

burglary and the crime drop: The need for spatial data signatures. *Journal of Criminal Justice*, 61, 90 – 100.

Hodgkinson, T., Andresen, M. A., and Farrell, G., 2016. The decline and locational shift of automotive theft: A local level analysis. *Journal of Criminal Justice*, 44(1), 49 - 57.

Malizia, N. (2013). The effect of data inaccuracy on tests of space-time interaction. *Transactions in GIS*, 17(3), 426–451.

Mazeika, D., and Summerton, D., 2017. The impact of geocoding method on the positional accuracy of residential burglaries reported to police. *Policing: An International Journal of Police Strategies and Management*, 40(2), 459-470,

Murray, A. Y., Grubesic, T. H., Wei, R., and Mack, E. A., 2011. A hybrid geocoding methodology for spatio-temporal data. *Transactions in GIS*, 15(6), 795 – 809.

O'Sullivan, D., & Unwin, D. J. (2010). *Geographic information analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Openshaw, S., 1984. Ecological fallacies and the analysis of areal census data. *Environment and Planning A*, 16(1), 17 – 31.

Ratcliffe, J. H., 2001. On the accuracy of TIGER type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science*, 15, 473 – 485.

Ratcliffe, J. H., 2004. Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*, 18, 61 – 72.

Roongpiboonsopit, D., and Karimi, H. A., 2010a. Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science, 24(7), 1081–1100*.

Roongpiboonsopit, D., and Karimi, H. A., 2010b. Quality assessment of online street and rooftop geocoding services. *Cartography and Geographic Information Science*, 37(4), 301-318.

Shah, T. I., Bell, S., and Wilson, K., 2014. Geocoding for public health research: Empirical comparison of two geocoding services applied to Canadian cities. *Canadian Geographer*, 58(4), 400–417.

Steenbeek, W., Vandeviver, C. Andresen, M. A., Malleson, N., and Wheeler, A., 2018. *sppt: Spatial Point Pattern Test*. R package version (0.1.5, 0.1.6). Retrieved from: https://github.com/wsteenbeek/sppt

Steenbeek, W., and Weisburd, D., 2016. Where the action is in crime? An examination of variability of crime across different spatial units in The Hague, 2001–2009. *Journal of Quantitative Criminology,* 32(3), 449-469.

Vandeviver, C., and Steenbeek, W., 2019. The (in)stability of residential burglary patterns on street segments: The case of Antwerp, Belgium 2005 - 2016. *Journal of Quantitative Criminology*, 35(1), 111 - 133.

Weisburd, D., 2015. The law of crime concentration and the criminology of place. *Criminology*, 53(2), 133 – 157.

Weisburd, D., Bushway, S., Lum, C., and Yang, S-M., 2004. Trajectories of crime at places: A longitudinal study of street segments in the City of Seattle. *Criminology*, 42(2), 283 – 321.

Weisburd, D., Groff, E. R., and Yang, S-M., 2012. *The criminology of place: Street segments and our understanding of the crime problem*. New York, NY: Oxford University Press.

Wheeler, A. P., Worden, R. E., and McLean, S. ., 2016. Replicating group-based trajectory

models of crime at micro-places in Albany, NY. *Journal of Quantitative Criminology*, 32(4), 589 – 612.

Wheeler, A. P., Steenbeek, W., and Andresen, M. A., 2018. Testing for similarity in area-based spatial patterns: Alternative methods to Andresen's spatial point pattern test. *Transactions in GIS*, 22(3), 760 - 774.

Whitsel, E. A., Quibrera, P. M., Smith, R. L., Catellier, D. J., Liao, D., Henley, A. C., and Heiss, G. (2006). Accuracy of commercial geocoding: assessment and implications. *Epidemiologic Perspectives & Innovations*, 3, Article 8.

Wu, X., and Lum, C., 2017. Measuring the spatial and temporal patterns of police proactivity. *Journal of Quantitative Criminology*, 33(4), 915 - 934.

Zandbergen, P. A., 2008. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, 32, 214 – 232.

Zimmerman, D. L., Fang, X., and Mazumdar, S. (2008). Spatial clustering of the failure to geocode and its implications for the detection of disease clustering. *Statistics in Medicine*, 27(1), 4254-4266.

Table 1. Counts of events (crime) and areal units of analysis

Type and number of events	Type and number of areal units
Vancouver, Canada (115 square kilometers)	
Theft of vehicle (1,474) Residential burglary (2,994) Theft (5,708) Theft from vehicle (12,809)	Census tracts (117) Dissemination areas (991) Street segments (18,445)
Leeds, England (552 square kilometers)	
Residential burglary (4,749) Shoplifting (5,666)	Super output areas (482) Output areas (2,543) Street segments (47,664)
The Hague, Netherlands (98 square kilometers)	
Theft of vehicle (1,025) Assault (2,478) Residential burglary (5,775) Street robbery (11,251)	Districts (44) Neighborhoods (114) Street segments (14,375)
Brisbane, Australia (1343 square kilometers)	
Graffiti (991) Assault (3,400) Residential burglary (5,327) Drugs (12,677)	Statistical area level 2 (137) Statistical area level 1 (2,707) Mesh blocks (14,150)
Antwerp, Belgium (205 square kilometers)	
Rape (980) Theft of vehicle (2,601) Assault (6267) Residential burglary (10,439)	Neighborhoods (44) Statistical sectors (307) Street segments (26,875)

Note. Mesh blocks are the smallest unit of geography available in the Australian national census; they are approximately the size of one city block.

Table 2. Geocoding match rate, *S*-Index = 0.90, crime type and area type

	Large area	Medium area	Small area
	Vancouver		
Theft of vehicle	30	80	90
Residential burglary	20	70	90
Theft	50	70	80
Theft from vehicle	10	45	85
	Leeds		
Residential burglary	35	75	90
Shoplifting	60	70	75
	The Hague		
Theft of vehicle	30	50	90
Assault	20	35	85
Residential burglary	15	15	85
Street robbery	10	25	85
	Brisbane		
Graffiti	55	80	85
Assault	30	80	85
Residential burglary	10	75	85
Drugs	10	65	80
	Antwerp		
Rape	30	70	90
Theft of vehicle	15	50	90
Assault	5	35	85
Residential burglary	5	25	80