Master Thesis

Research Master in Spatial Science

Alessandro Fois *S4115937*

Supervisor: Dr. Daniella Vos
Second Supervisor: Dr. Michiel Daams

Faculty of Spatial Sciences
University of Groningen

Empirical approach to the problem of point pattern analysis in case of geocoded data

"For many geographers, point pattern analysis will conjure up images of 'nearest-neighbour analysis' applied inappropriately to data sets of doubtful relevance." - Gatrell, Bailey, Diggle, Rowlingson (1996)

Abstract (max 100): 100

This research hypothesizes that the geocoding process represents a transformation of space from continuous to discrete invalidating the homogeneity assumption of the Poisson point process. Therefore, we propose to adapt the existing methods using the whole spatial database of addresses to calculate the expected value and test the null hypothesis of complete spatial randomness. The results show inconsistency in the ordinary methods when a true random pattern is tested while the proposed method is able to distinguish a random pattern from a non random one. We concluded that the geocoding introduces a bias depending on the field and method used.

Introduction

The geocoding process consists in the transformation of descriptive information, typically an address, into geographic coordinates (Goldberg et al., 2007; Boscoe, 2008; Zandbergen, 2008). Through this procedure it is possible to return a precise spatial dimension to numerous information that otherwise would have to be examined by being grouped in a container, e.g. the administrative boundary, which dilutes the spatial dynamics of the observed phenomenon on a wider scale. Consequently, the procedure has become an essential step in many disciplines that share the same need for detailed spatial information on a phenomenon. Epidemiology, criminology and economics among others disciplines make extensive use of geocoding (See Gatrell et al., 1996; Anselin et al., 2000; Bonner et al., 2003; Ratcliffe, 2004, Duranton and Overman, 2005; Hart and Zandbergen, 2013; Owusu et al., 2017, p.42).

Although geocoding has made possible the use of spatial statistics in many fields with difficulty in mapping the data, there are some common and widely discussed limitations, mainly concerning the geometrical aspects such as location inaccuracy, imprecision and incompleteness (e.g. Malizia, 2013) but also involving theoretical aspects depending on the phenomenon observed. Although these types of errors are inherent to any positioning method, geocoding adds uncertainty (to an already complex system) especially if the process passes through spatial databases that are inaccessible to the user. Despite uncertainty is commonly used in geography to refer all the sources of attribute and positional errors that lead the dataset to deviate from reality (Goodchild, 1998; Brown and Heuvelink, 2008; Züfle, 2021), in this context it can be used to address the impossibility to

verify the actual contents of the spatial database. This is consistent with a broader definition given by Goodchild (2002, p5) which clarifies how uncertainty can be a '*measure of the difference between the data and the meaning attached to the data by the current user*'.

The accuracy of the locations and the match rate is a limitation inerent the geocoding process, irrespective of which spatial database is used, commercial or public (Whitselet al., 2006; Zandbergen, 2009). This is in contrast to the common and misplaced trust in the geocoding process as a tool systematically capable of obtaining accurate data mapping (Malizia, 2013). Consequently, all the possible phenomena are affected by accuracy errors once transformed into point patterns through geocoding. Moreover, the accuracy of the location excluding the internal source of errors, is related to the quality of the spatial database from which the addresses are excerpted (Goldberg et al., 2007). The process of associating the address in a location passes through different levels of accuracy; area, street, and house numbers, the matching rate of these geometrical levels gives the accuracy, and all the points must eventually have the same rate (Owunsu et al., 2017, p.47; Sahar et al., 2019). Therefore, even if the final accuracy of the point location results in geographic error measurable in spatial coordinates distances, the source and amount of the error remain unknown, unless it is compared with a measurement on the field (Owunsu et al., 2017, p.50). Furthermore, the error can be completely independent from the quality of the information collected about the observed phenomenon. A mismatch between the text of the recorded address and its equivalent in the spatial database will lead to a location error as well as a correct matching of the addresses with inaccuracy in the spatial database.

In addition to the sources of uncertainty due to the geocoding process used to map the point pattern, there are also limitations due to statistics. Depending on the combination of the statistical method used and the degree of accuracy of the geocoding, the results can be very different (De Luca and Kanaroglou, 2008). When dealing with point patterns should always be considered a possible edge effect (See Ingram, 1978; Cressie, 1991, p610; Brix and Kendall, 2002; Yamada and Rogerson, 2003) to which magnitude depends on the statistical method used and the phenomenon observed (Baddeley et al., 2015, p212-216). The surface effect, because the intensity of the point pattern is directly related with the area (Ingram, 1978). The surface can also interfere with the geocoding process, making a point fall in wrong administrative boundaries (Goldberg and Cockburn, 2012).

There exists a second group of limitations, namely, ontological limitations relating to the observed phenomena. Discussing all of them is not possible, neither is the purpose of this work, while it is relevant reflecting on their existence and how they can change depending on the observed phenomenon. For instance, according to Gatrell et al. (1996), the assumption of epidemiological events, represented by the address of residence of the infected, imply that people do not interact with the environment outside their residence even though we know they could have been exposed elsewhere. While criminology, according to the place based and routine theories, assumes that most of the events originate around very small areas compared to the observed one (Anselin et al., 2000).

Consequently, in both fields, the points can represent a set of events, approximating the phenomenon that causes them and, in both cases, the real location of the source is unknown.

Nevertheless, if we consider all the inevitables error and limitation affecting the point pattern a tolerable approximation, testing the hypothesis against complete spatial randomness (CSR) it is a reasonable beginning (Cressie, 1991; Kulldorff, 1998; Anselin et al., 2000; Duranton and Overman, 2005). The reason why it is important to discern between a random pattern and a non-random one, originates from the need to understand if there are explanations that justify the disposition of the pattern (Aldstadt, 2010). Clearly, the concept of randomness is decisive, which is usually defined by the probability distribution associated with the phenomenon and its relative spatial point process (Daley and Vere-Jones, 2003, p19). However, when the probability distribution is too complex, it may be advantageous to calculate it empirically using Monte Carlo, to then compare the values under conditions of randomness (Cressie, 1991, p635; Hope, 1968). However, it is often neglected that Monte Carlo is a computer-generated experiment and can be affected by the hardware and software limitations that define the pseudo-random generator, which should be considered as possible sources of errors (Van Niel and Laffan, 2003). Therefore, the computational limits of the pseudo random generator should be checked during the analysis to avoid further sources of errors.

A point pattern generated through the geocoding procedure has, therefore, numerous specific limitations to be accounted as potential source of errors. While these listed limitations are extensively discussed in the literature, the role of the preexisting structure due to geocoding remains unexplored. In other words, the effects of the predetermined point structure inherited by the spatial database on spatial randomness test remains unclear, even considering all the known limitations associated with geocoding. It is also noted that the development of methods of analysis and hypothesis testing for point patterns are chronologically prior to the diffusion of reliable and cheap geocoding methods. According to Goldberg et al. (2007), the first geocoding systems began to be developed in the 60s and the usage of precisions higher than postal codes in the 90s. As already observed, the structure of a spatial database for geocoding complies with cartographic and computer-based rules for the allocation of geometries (Goldberg et al., 2007). These rules are foreign to the disciplines of epidemiology, criminology or economics, nevertheless they can find an acceptable approximation in the locations identified by the geocoded points. For all of these reasons, in this research, we ask if the geocoding process used to map a pattern introduces a bias in the point pattern analysis.
Does the geocoding process violate the assumptions of the (in)homogeneous poisson process?

Theoretical Framework

The search for patterns within spatially distributed data has a long tradition in disciplines with a geographic approach. A widespread opinion is that variables correlation over space

depends on their distance. This is frequently synthesized with Tobler's first law (1970). Nevertheless, according to Haining (2010), the Tobler's first law states a concept which, although it was not defined in terms of law, was already discussed in the early twentieth century by Student and Fisher. The concept of intrinsic relationship is pivotal in geography because it implies spatial autocorrelation; thus, a non-random arrangement of objects in space, making CSR a special case (Goodchild, 2010; Getis, 2010a). Therefore, tests of CSR can be seen as meaningless when all the implicit variables are known (Gatrell et al., 1996; Aldstadt, 2010). However, the CSR test remains a relevant starting point (Illian et al., 2008, p58) and methods based on distances and the homogeneous poisson process are still very relevant in epidemiology (see Auchincloss et al., 2012; Kirby et al., 2017). Therefore, the degree of knowledge of the observed phenomenon and the variables involved are paramount in deciding the most appropriate conceptualization and analysis model to be applied.

According to Cressie (1991, p8, 52), there are three kinds of possible abstractions for the spatial data: 'continuous spatial index', 'lattice index', and 'point pattern index'; which he summarizes in the following way:

$\{\mathbf{Z}(\mathbf{s}): s \in \mathbf{D}\}$

Where **s** is a vector $\in$ **R** representing a location in a Euclidean space; the function **Z(s)** is the data observed at the given $\mathbf{s_n}$ coordinates; **D** is the index; a probability space within **s** varies randomly, therefore, an observed pattern can also be seen as one of the possible random realizations of the space defined by **Z** and **D**.
Starting from this simple definition, we can address all the possible models, where geostatistics mainly cover the continuous spatial index while the lattice applies to models that foresee the use of raster data and polygonal spatial units. Therefore, the point pattern is seen as a specific case where the space is a subset of **R** and the topology has a countable base (Cressie, 1991, p619). However, the generic definition does not exclude the possibility to remodulate the point patterns in the same manner as is made for the lattice data. Given the definition of the geocoding process described by Goldberg et al. (2007), and the fixed structure of the spatial database where the addresses are stored, the definition of lattice data seems to be more coherent. Applied to the mathematical definition, it means taking D as a countable collection of spatial sites that are, in this case, points representing the locations where the event can occur (Cressie, 1991, p383).

Many methods for point pattern analysis are tested against the CSR hypothesis (Diggle, 2013, p99), this is equivalent to assume they follow a homogeneous poisson process (Ripley, 1977; Illian et al., 2008, p66), although the homogeneity requirement is not always specified (Cressie, 1991, p586). Consequently, if the observed model has a distribution too far from the Poisson probability model, then it is not randomly distributed and could be clustered or dispersed. However, if we use the point pattern index models as they are defined, we are forced to hypothesize, and simulate, a homogeneous poisson process over the study region surface, in contrast to the limits introduced by the geocoding process.

Alternatively, the complete list of points allocated in the spatial database used for the geocoding process can be considered as an abstraction of the space, and it can be described with the lattice index defined by Cressie (1991). In other words, the list of locations recorded in the spatial database represents the sample space of all the possible outcomes and the point pattern is a (random) subset of that sample space. Thus, the sample space has a discrete, countable finite or infinite number of elements depending on whether the observed phenomenon can have a finite or infinite number of events.

Since the lattice index defined by Cressie (1991, p383-384) was applied to polygonal spatial units or raster, where, for instance, the county seat coordinates could represent the nodes of the lattice, its application to the geocoding point grid is straightforward. However, to extend the concept of lattice to existing point pattern models it is appropriate to introduce changes in these models in a way that is consistent with the definition of the sample space. Changes that, most likely, can be extended in a similar way to different models of point patterns when based on the same assumptions. Indeed, the point pattern methods can be distinguished in two main groups, the distance-based methods or techniques (e.g. Nearest-Neighbor Distribution Functions and the K function) and those on the area (e.g. Kernel Estimators of the Intensity Function) (Gatrell et al., 1996;). For example, in case the model considers the area, if the sample space is represented by the complete set of points, the intensity of an event will not be the number of events per unit of surface area but the number of events that occurred for a group of possible points. However, this involves the introduction of a criterion, and a relative function, which keeps the number of points from the set almost constant. Therefore adapting models based solely on distance is simpler. Although distance-based methods are not the most used among disciplines that make extensive use of geocoded data, they are generally easy to adapt.

The nearest neighbor methods are some of the most used (Ingram, 1978), and within this group, the Clark and Evans (1954) statistic represent an early and relatively easy approach. Originally it was designed to identify the presence of non-random patterns, clustered or dispersed, in plant populations. Nevertheless, it has also been used outside its primary field, probably due to its simplicity and its implementation in the well-known software ArcGis (e.g. Lee and Wong, 2001, p72; Scott and Janikas, 2010). The Clark-Evans test requires only a point pattern and the surface in which events can occur.

This method of analysis is, in fact, a hypothesis test of CSR which indicates whether the pattern is random, clustered or dispersed, with respect to the assumptions made in the model. However, the Clark-Evans method is uninformative because first, it has certain intrinsic limitations and second, it does not identify in which portion of the pattern the cluster is located. Given these reasons, this method has to be followed by a further analysis (Aldstadt, 2010). In spite of the stated problems and limitations the lattice index concept can be applied to the Clark-Evans test due to its simplicity. The key of the modifications is performing the test without accounting the area and calculating the expected value with

Monte Carlo as suggested by Baddeley et al. (2015, p258) and Diggle (2013, p24-25) as explained in the methodology.

Methodology

In order to verify the hypothesis that the geocoding procedure introduces a transformation of the sample space, from continuous to discrete (lattice index), was selected a published work that makes use of the Clark-Evans' nearest neighbor method. The latter test was replicated, adjusting the test in order to be consistent with the hypothesis of lattice sample space, and the results were compared. Moreover, the methodology and its proposed variation have been compared using some random point pattern to test their capability to recognize a random pattern from a geocoded dataset. Although Clark-Evans' nearest neighbor method was designed for use in the ecological context, it is sometimes applied in other fields. The usage of the method outside its design specifications increases the chances of violating its assumptions, in particular of homogeneity, that according to Getis it happens quite often (2010b). However, the use of the Monte Carlo method to calculate the expected value in a lattice structure, extends the possible contexts of applications of the test because the sample space corresponds to the points where events can occur.

The work of Enthoven and Brouwer (2019) analyzes the spatial dynamics of sustainable restaurants with various methods of which the Clark-Evans' nearest neighbor is the first approach to the analysis, used as a CSR test applied to the point pattern distribution generated through a geocoding process. The subject of the paper, sustainable restaurants, is irrelevant to the research question considered in the present paper. They can be addressed simply as events, ignoring what they represent except that they are observations from a point pattern inhomogeneously distributed. What is relevant is to verify the hypothesis by comparing the interaction between the points under the assumption of continuous space, to the points under the assumption of discrete space. As a consequence of the widespread use of the geocoding process, this method can be applied to many fields. The scholar can evaluate the conditions in each case.

To verify the hypothesis using Clark-Evans' test, some simple changes are necessary. These changes are also necessary for cases where the expected value is calculated with Monte Carlo. The difference lies precisely in the sample space; the original method was designed to approximate the behavior of plants, so the expected value was defined under the condition of homogeneity, over the entire surface, even with Monte Carlo. Consistently with the hypothesis, however, the expected value for the proposed method is calculated with respect to all the possible points that compose the database used for geocoding. This is equivalent to assuming a stationary points process since it is like a frame only, containing all the points.

The Clark-Evans test (1954), can be expressed as follows:

$$R = \frac{r_A}{r_E} = \frac{\frac{\Sigma r}{N}}{\frac{1}{2\sqrt{\rho}}}$$

Where r_A is the mean nearest distance of the observed pattern, and r_E is the analytical mean distance expected under condition of CSR. N is the number of events and ϱ is the density for the given area in square meters. The index R can assume values between the interval $0 \leq R \leq 2.1491$, where R=1 if the pattern is randomly distributed, R=0 in case all the points are overlapped and reach its maximum value when the events are distributed in a even hexagonal pattern (see Clark and Evans, 1954).

This test assumes a homogeneous distribution, thus the population distribution in a country violates this assumption. The strategy used to solve this problem is calculating the empirical distribution function through a Monte Carlo simulation accounting for the population density. This distribution function was obtained by generating 999 patterns with N events located at random coordinates inside the country boundaries and calculating the average distance (in the same manner of the observed pattern). Therefore, in this case, the density parameter is given by the population density because it acts as a probability factor. Notably, the distribution of mean distances in case of CSR is Gaussian (Diggle, 2013, p25).

The expected value for the proposed method was obtained by calculating the mean distance of 999 sets of N events with locations randomly chosen from the spatial database used for the geocoding process. Therefore, this method required re-geocoding the case study dataset using a freely accessible spatial database. While the first Monte Carlo simulation was obtained with the tool provided by the *spatstat* library of *R* (Baddeley et al., 2015, p259), the second simulation for the proposed method was written to run in R as well, following line of the Baddely code.

The analysis was then divided into two sections, first the comparison between the results of the case study obtained with the classical method and the one proposed, second, the comparison of the results of the two methods with respect to some random patterns.

Preliminarily, it was verified that both the dataset and the software used were able to generate the same results published in the Enthoven and Brouwer paper. Consequently, the first step was to replicate the Clark-Evans test in its classic form as performed in the document. Furthermore, as part of this preliminary analysis, the test was repeated using Monte Carlo to calculate the expected value for the entire surface, accounting for the population density. In doing so, the homogeneity hypothesis was not violated. Incidentally, the comparison between the result obtained with Monte Carlo and the value obtained with the analytic function provides a measure of the difference for this case study. This was necessary to be able to compare the results with the same procedure applied to the lattice index, the last test for this section. The spatial database of the addresses is a good approximation of the population density, although it does not contain information on

household components. Therefore, the only remaining difference between the expected value calculated on the continuous surface and the one calculated with respect to the lattice structure is precisely the presence of the area. Eventually, comparing these two results will provide insight into the effects of geocoding on the model assumptions.

The second section of analysis consists in verifying the accuracy and reliability of the model, again by comparing the results obtained by the ordinary method with the one proposed. This is a common approach incidentally used by Clark and Evans (1954) to empirically verify the expected values in case of randomness. Both the models are tested with some random distributions. In this case, the aim is to compare the capability of the models in recognizing a complete random pattern.

Dataset and software

The data for this study was composed by a point pattern as a case study, a spatial database containing the addresses in the Netherlands, the population density distribution and administrative boundaries of the Netherlands.

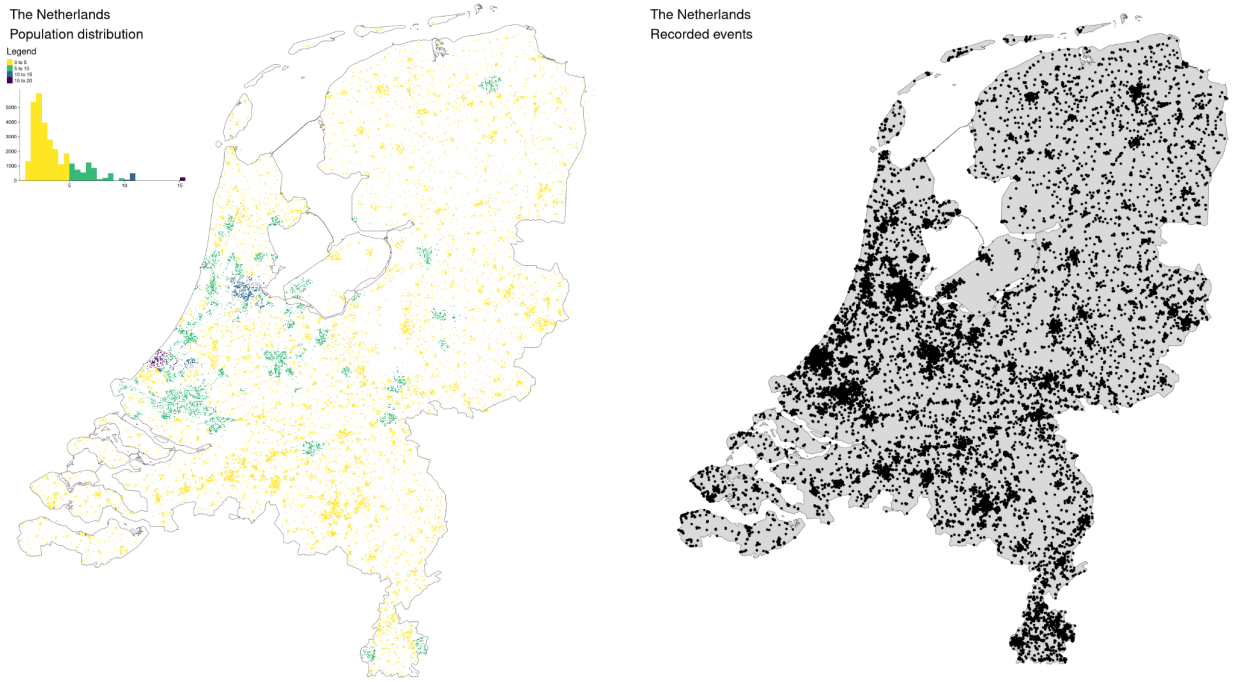

Figure 1.

The case study consisted in the same dataset used in the article of Enthoven and Brouwer. The dataset was provided by the corresponding author M. Enthoven and contained the list of sustainable restaurants in the Netherlands updated until the year 2013. The dataset was structured in two files, one containing a list of addresses (id, address, coordinates) and a GIS file containing the points geocoded for their article.

The population density of the Netherlands was retrieved from The WorldPop project of the University of Southampton, UK (WorldPop, 2021)

The administrative boundaries and the list of addresses of the Netherlands is freely accessible from PDOK, the open data portal for the geographical data of the Netherlands (PDOK, 2021).
All software processing was performed using the statistical software R (R Core Team, 2021). While the data was converted from the original format used by ArcGIS and imported into a PostgreSQL / PostGIS database

Results

The first section of the research analyzes the case study (CS henceforth) by first replicating the procedure followed by Enthoven and Brouwer, then empirically calculating the expected value in conditions of CSR with Monte Carlo. The results of the latter are then compared with the expected value obtained empirically with Monte Carlo but consistently with the hypothesis of discrete space proposed.

First section, the case study

A summary of the results can be seen in table 1. The results obtained in the original paper have been successfully replicated. This confirms that the data is the same and the software implements the same algorithm, despite the fact that the CS used ArcGIS while R was chosen in this work. The table also shows the column 'adjusted' which refers to a corrected version of the dataset. To be able to apply the proposed methodology, it was necessary to replicate the geocoding procedure using a freely accessible database.

Clark-Evans ANN analysis	**original dataset***	**adjusted dataset****	**original dataset*****
Observed Mean Distance (r_A)	173.7942 Meters	178.6267 Meters	173.7942 Meters
Expected Mean Distance (r_E)	917.4406 Meters	643.8703 Meters	1045.921 Meters
Nearest Neighbor Ratio (R)	0.189434	0.277427	0.1661639
z-score	-305.493797	-268.621258	-314.264
p-value	0.000000	0.000000	0.000000
number of events	38812	37762	38812
Study Area	130'672 Km^2	62'620 Km^2	169'834 Km^2

*The values are referred to the output of ArcMap 10.5.3 used to replicate the CS in the same manner

** Before the re-geocoding process

***The same ANN analysis on the original dataset performed with R. The difference is due to the computation of the minimal surface that differs in R. With the proper corrections the values collimates

Table 1.

During this procedure of re-geocoding, a group of 1050 points appear to be misplaced, far away from the Netherlands. Typically, the geocoding softwares accumulate the output of mismatching entries at the origin of the axes or in a largely approximated location. This likely happened in the original dataset and resulted in deleting the points, after having verified that it was not possible to find the appropriate location. Consequently, the observed average distance is increased, because the points were stacked, and the surface occupied by the points was reduced at the Netherlands bounding box. Evidently, by using this adjusted dataset, the point pattern resulted with a higher R index (0.28) but still significantly clustered. Furthermore, the expected value obtained from the analytical calculation is in this case reduced to 643.9. The latter result is mainly due to the variation of the surface. However, it remains overestimated as seen from the comparison in Figure 2. By calculating the expected value under conditions of CSR with Monte Carlo, which takes into account both the population and the administrative boundaries, the average value drops to 303.6, as can be seen from the summary in table 2.

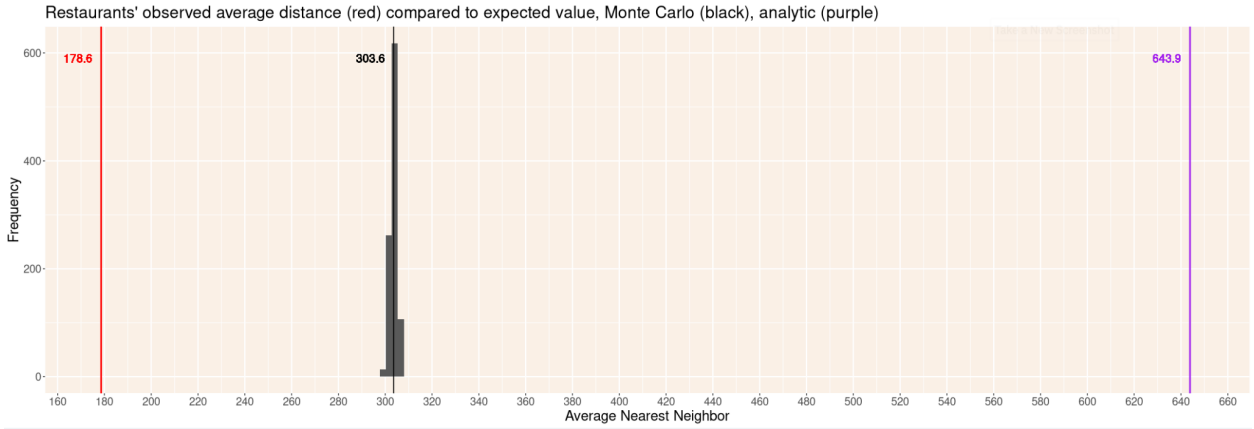

Figure 2.

Therefore, although the clustering condition of the point pattern is confirmed, the value of the R index rises to 0.59 which means, according to Clark and Evans (1954), the nearest neighbours are on average 59% distributed, as expected under condition of randomness.

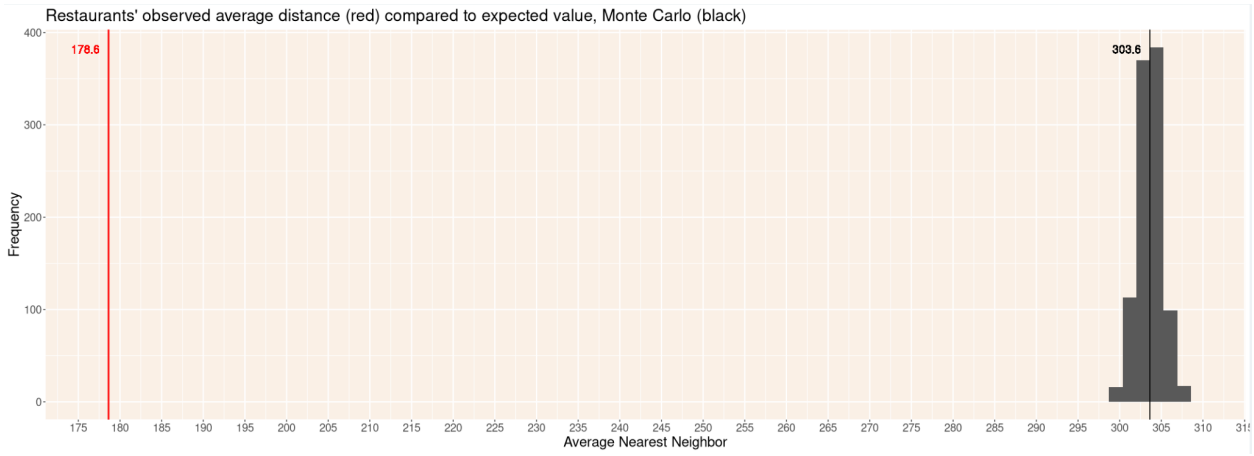

Figure 3.

Nevertheless, even with an index R 0.59, we could be led to consider the clustered pattern, as shown in Figure 2.

Monte Carlo ANN analysis	**classic method***	**proposed method****
Observed Mean Distance (r_A)	178.6267 Meters	189.4021 Meters
Expected Mean Distance (r_E)	303.6 Meters	246.1 Meters
Nearest Neighbor Ratio (R)	0.588311	0.769614
pseudo p-value	0.001	0.001
number of events	37762	37762

* adjusted dataset before re-geocoding | ** adjusted dataset after re-geocoding

Table 2.

However, when the proposed method is applied, the observed mean distance changes. This is due to the new geocoding process which uses a different spatial database. The value moves from the original 178.6 of the adjusted pattern to 189.4. Notice that the change in the spatial database has caused the R index to become 0.62 for the classic Monte Carlo procedure and 0.77 using the proposed one. It is interesting to note that the variation of the geocoding database has simultaneously produced an increase in the observed average distance and a decrease in the expected average distance which is now 246.1 (see figure 4 and table 2) for the proposed method and remains the same for the classic Monte Carlo. Therefore, the deviation from the mean distance in the case of CSR is no longer so marked, despite that the observed pattern is slightly clustered in this analysis.

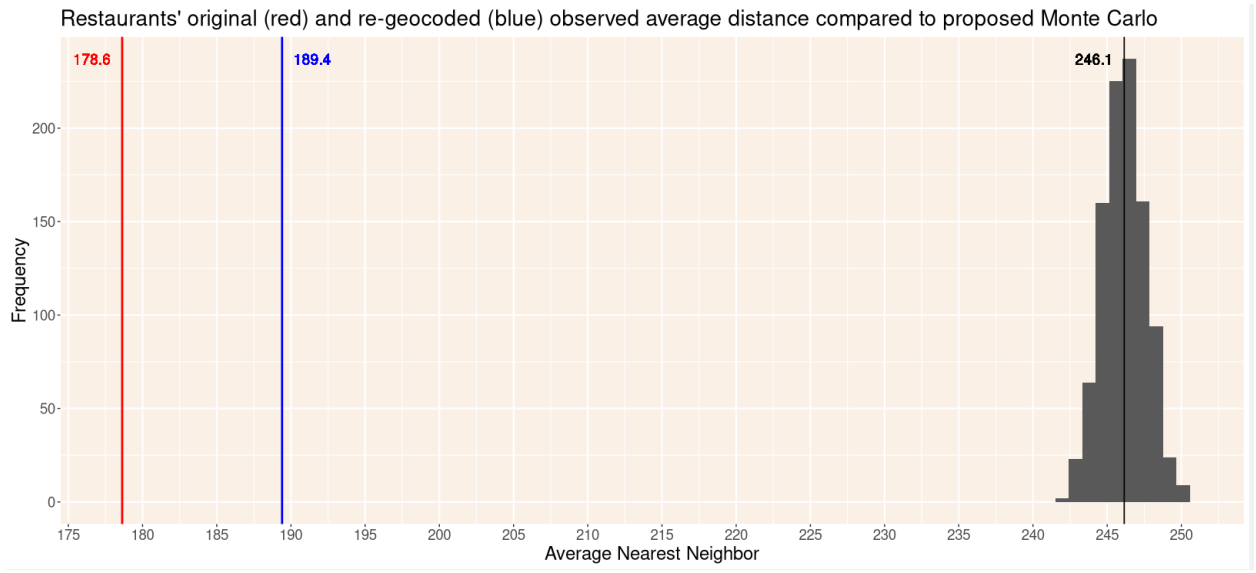

Figure 4.

The application of the lattice concept meant limiting the possible values of the random pattern to the same values the observed event can obtain when geocoded. Moreover, the pseudo p-value was 0.001, thus, given these features, the results should be considered more reliable than the previous one. It is important to add that the observed pattern was treated as a generic pattern, *id est*, it was assumed that each address of the spatial database was equally possible. However, we know that the CS is about sustainable restaurants and, as will be seen below, the expected distance value for a business subset decreases further. Nevertheless, it was considered wrong to use this subset in comparison with the CS, since the latter does not make this distinction. Furthermore, it is believed that these reflections belong to economists and should be supported by the literature of the field for contextualizing the chosen models.

Second section, the random pattern

The proposed method seems to have introduced an increase in the quality of the measurement so far. This section will show the results obtained measuring various random points patterns, generated by randomly selecting N points from the spatial database of the addresses.This is necessary to determine whether the results of the proposed method are a circumstantial fact or more closely correlated with the hypotheses described in the methodology.

Only two simulations are presented below for the sake of brevity, one sampled from the complete list, and one sampled from the subset of business activities.

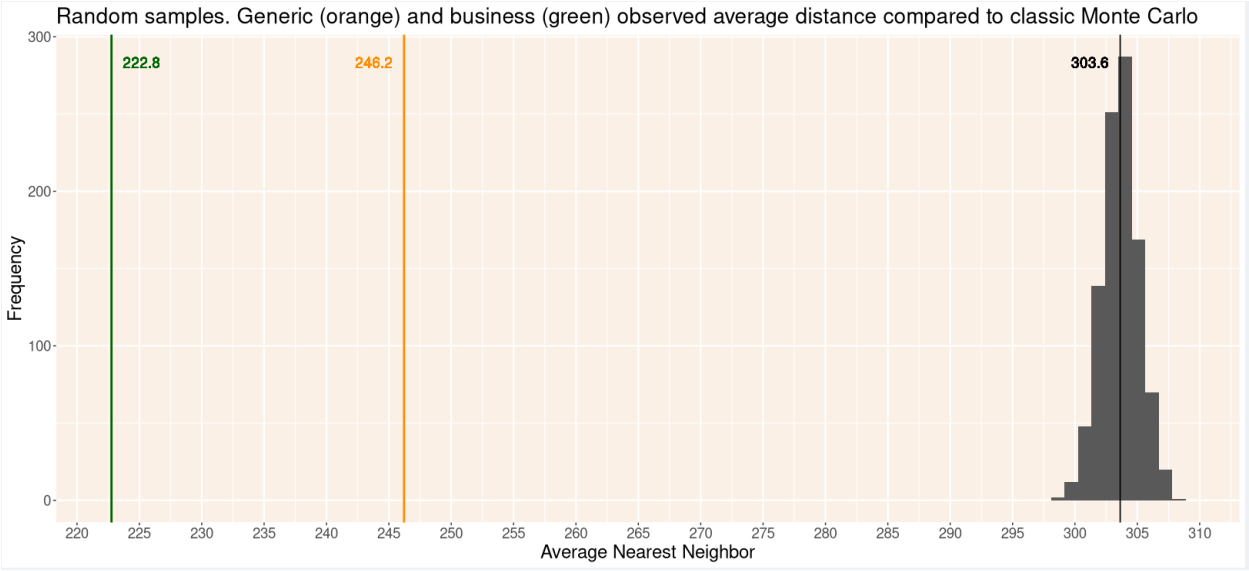

Figure 5.

As presented in the previous section, the random pattern consists of a sample of the same number of events as the CS. The reason lies in facilitating the interpretation and creating a realistic simulation of a possible phenomenon under the condition of randomness.

The results of the tests on the random samples are summarized in table 3. Clearly, the expected value generated empirically with Monte Carlo in a classical way does not change either with respect to the previous tests or with respect to the business subset, except for very small values. The small difference between the values is because of the high number of simulations. While to adapt the expected value to the business subset was needed data about their density, which unfortunately was not found.

Figure 5 shows the comparison between the classic Monte Carlo method and the random patterns. The orange line represents the observed average nearest neighbor of a random sample that is a realization of all possible addresses; its value is 246.2 meters. The green line represents the observed average nearest neighbor of a random sample that is a realization of the subset of addresses corresponding to commercial activities and firms; its value is 222.8 meters.

n=37662	**generic random sample**	**business random sample**
Observed Mean Distance (r_A)	246.2204 Meters	222.7652 Meters
Expected Mean Dist. -Classic MC- (r_E)	303.6 Meters	303.6 Meters*
Expected Mean Dist. -Proposed MC- (r_E)	246.3 Meters	221.9 Meters
Expected Mean Dist. -Analytic- (r_E)	740.6156 Meters	740.6352 Meters

NN Ratio -Analytic- (R)	0.332445	0.300784
NN Ratio -Classic Monte Carlo- (R)	0.811003	0.733746
NN Ratio -Proposed Monte Carlo- (R)	0.999679	1.003899
z-score (analytic only)	-247.839	-259.601
pseudo p-value	0.001	0.001

* The firm distribution should have been calculated using the firm density instead of the population

Table 3.

Clearly, applying the analytical method in this context is inappropriate. The reason is, we are observing addresses, mainly composed of dwellings and the addresses are not homogeneously distributed. In this case, with index R values around 0.3, we would be forced to reject the null hypothesis, but we know that the pattern is random. However, when the random patterns are tested using the classic Monte Carlo method, considering the population density, the R index is 0.81 for the generic sample and 0.73 for the firms, thus, blandly clustered. The decision as to whether or not to reject the null hypothesis should be evaluated according to the context and the literature of the field of study. Perhaps we can test the significance with the appropriate function. This statistic does not exclude that a non-random pattern is distributed as expected under the condition of CSR (Clark and Evans, 1954). Notably, in this case, it is not possible to generate an empirical distribution with the Monte Carlo of the business subset, because a database of the firms' density cannot be found. Therefore, this represents a further limitation of this classic method.

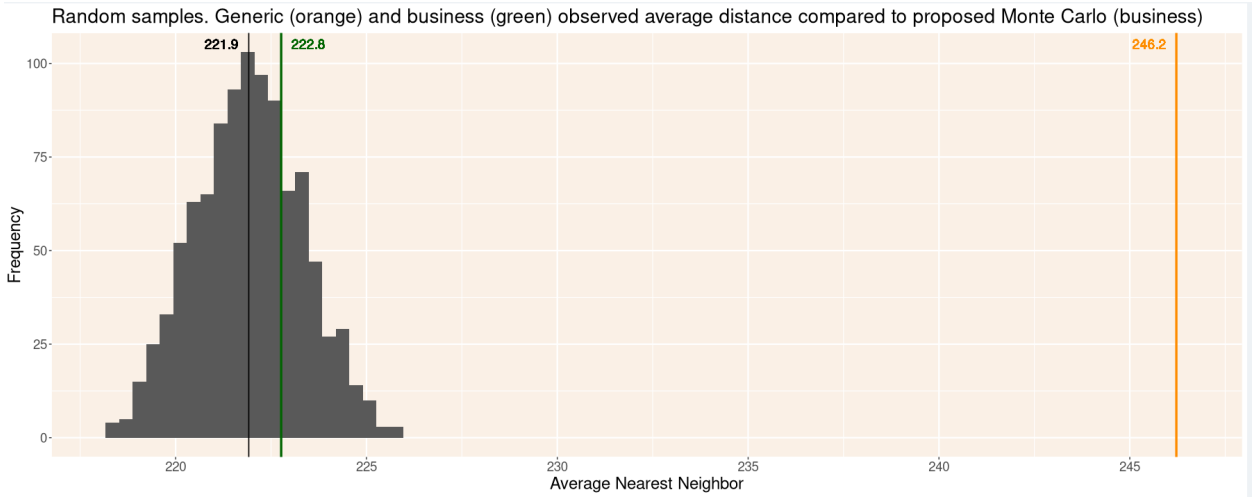

Figure 6.

However, as can be seen from Figures 6 and 7, the proposed method allows us to evaluate the distribution reliably. The R index is approximately 1 for both firms and a generic

sample. This is not surprising given that the observed random sample is a single representation of the same dataset used to generate the empirical distribution with Monte Carlo. Compared to the classic Monte Carlo method, here it is possible to distinguish between the subsets of points because it is an attribute stored in the spatial database.

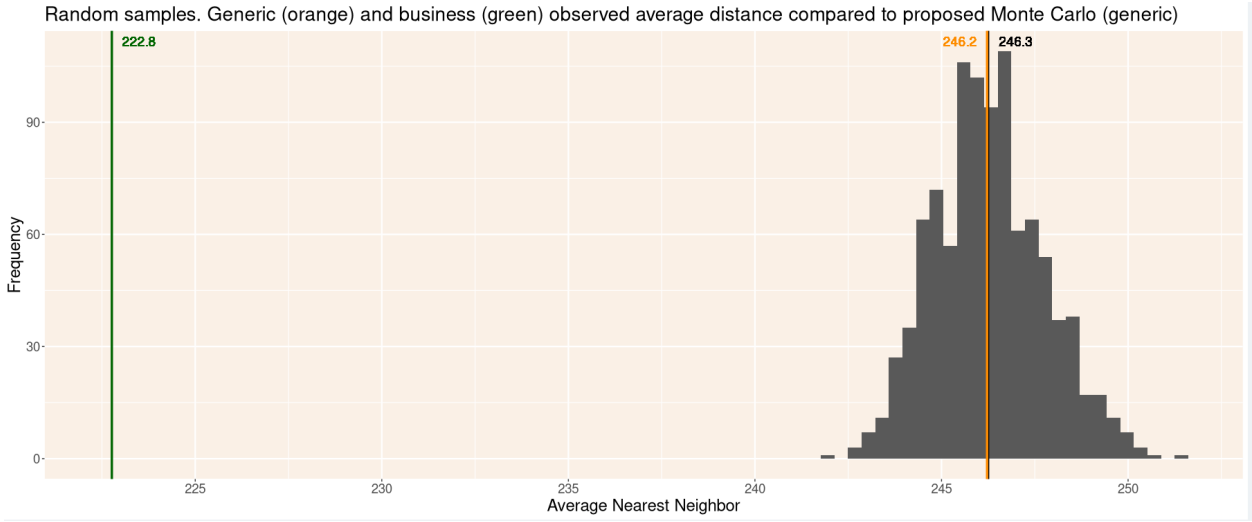

Figure 7.

This is a further insight for evaluating the hypothesis that the geocoding plays an important role, the value of 222.8 appears to be slightly clustered compared to the expected value for a generic distribution, while the value of 246.2 seems to be dispersed if it is compared with the inappropriate subset.

Discussion

This research aimed to investigate whether the geocoding process represents a transformation of the sample space. According to the results obtained from these preliminary experiments, it can be stated that the mapping of a pattern using an extract of a pre-established point lattice introduces a further limitation in the classical point pattern analysis models. The comparison between the classical method and the proposed one, which considers the spatial database of geocoding as a new sample space, gives a first measure of the degree of approximation given by the surface on which the events are observed, and on which they cannot move except in the established positions in contrast with the assumption of space continuity. Even if the space is not assumed to be homogeneous (for instance due to the population density) it is still considered continuous. Furthermore, the classic Monte Carlo method represents a significant increase in quality

compared to the analytical method for determining the expected value, when the Clark-Evans test is applied to a non-environmental context. The homogeneity assumption in an environmental space is indeed reasonable. Nevertheless, the introduced precision improvement is not sufficient to exclude the presence of clustered or dispersed patterns. A researcher who applied the method as a preliminary step in the absence of a priori information on the distribution, would be induced to further analyze the pattern using other methods to identify the position of the cluster, ignoring that on the global scale, there was no difference compared to a random distribution. This is an important detail considering that a pattern of points distributed in space according to a non-random criterion could be arranged similarly to a completely random pattern and *vice versa*. Even a small uncertainty factor as emerged from the results of this study, could easily induce a statistical error of both types I and II. Supporting the choice of rejecting the null hypothesis on the basis of the literature of the field is much more difficult when the value of r_A is almost equal to r_E.

Moreover, it is worth reflecting on the frequent practice of mixing geocoding databases in order to get the best match. The result of this work, although further research is needed, shows how important it is to have access to the database that generates the geocoding. A closed spatial database acts as a black box adding uncertainty to models that already have numerous approximations. Furthermore, mixing the geocoding sources, with respect to the issue discussed here, only increases the difficulty of tracing the transformation introduced. In addition, even if the geocoding datasets used were all freely accessible, there would remain the uncertainty of how to implement the test methodology. From this point of view, a point map geocoded with more than one independent spatial database would represent an introduction of a systematic error in all the locations that are different from the spatial database chosen as a new sample space reference. However, a possible solution would be to merge the two or more spatial databases before starting the geocoding procedure so the data integrate with each other and the final map would result, geocoded from one single spatial database. However, this would be a complex procedure that requires advanced GIS skills.

Furthermore, it is not possible to easily access reliable open data in most geographical areas of the world, therefore a realistic and recommendable solution would be to apply a (classic) Monte Carlo method, considering the distribution of the population as closely as possible to reality, or the relevant probability distribution is followed by the observed pattern. Noticing though that, an empirical distribution function of an inhomogeneous Poisson process, calculated in such a way, would account for locations outside the sample space while the proposed one did not. One example of the grade of the approximation can be the difference between the output of the two methods for the CS. It is hard to determine whether the distance range obtained in this research can be generalized to other observations without further research. Nevertheless, it would be advisable to consider only extreme values of the R index grouped or dispersed and apply all the appropriate

evaluations case by case.

Conclusion

In conclusion, the answer to the research question whether or not the geocoding process introduces a bias in the point pattern analysis is that it certainly does. This is demonstrated by the difference in the results between the calculation of the expected value under conditions of randomness, obtained by the classic Monte Carlo method and the proposed one. Moreover, a measure of the bias between the analytical method and both of the Monte Carlo methods emerged as well. The former assumes the homogeneous distribution which is unrealistic therefore it generates bias. In fact, it is believed that the bias detected between the two empirical distributions, one generated from the spatial database used for geocoding and one generated by stochastic events over the space, represents a further variation in the distribution of events in space. In other words, it would be equivalent to a transformation of the sample space which is reduced to a discrete subset of the real space. Eventually, a classic Monte Carlo simulation must account also for the approximation given by the difference between the continuous and discrete sample space in order to reduce the bias.

A more general point that emerged concerns the theory. Most of these methods were defined many years before the geocoding process became easily accessible, fairly reliable, and available almost everywhere. The literature review showed that the general theory synthesized mainly by Cressie (1991) and Diggle (2013) is still the current one, however, geocoding is not mentioned in their work. Therefore, it seems that it is possible to analyze the question posed in this work from a purely theoretical point of view and that this could be sufficient to clarify whether the geocoding process is actually a transformation of the sample space as assumed here. Regarding the method chosen to test the hypothesis, the Clark-Evans test, its use has often been criticized concerning its conceptualization of distance, which in the original model is linear but in many applications it is not. Frequently, the events that characterize the city relate to each other through the road network and in this field many authors have introduced remarkable variants. However, there are phenomena that occur in the urban network but related one to the other as the crow flies, for example the case of many diseases transmitted through mosquitoes that act as vectors. In this case, it would not make sense to apply the linear network variants of the point pattern analysis, but evaluations such as those proposed in this paper should be considered. The existence of a bias in the Clark-Evans test when the pattern is geocoded, suggests that a similar bias can also be detected in other distance-based methods, such as K-function or Knox space-time. Furthermore, geocoding introduces bias also in other methods based on spatial units since the spatial database of addresses intrinsically represents a distribution of the population. Therefore, further research is certainly recommended for both the distance based methods and the area based methods.

Limitations of this research

There are some important limitations in this research. The most important is probably the measurement scale. Each spatial database containing the addresses for geocoding is built with a very high level of detail, capable of discerning individual dwellings. Assuming that an observed phenomenon is geocoded to the maximum level of detail available, it follows that that phenomenon also has the same scale of the spatial database, whether the information in the researcher's possession was of greater or lower detail. This could be a negligible problem if all the data used in the analyses have the same level of detail. However, when a dataset in the analysis has a less accurate scale than the other data, that dataset with lower detail should become the reference scale, otherwise this could introduce a bias. As we have seen, this research is affected by this limitation as well. The population distribution could have influenced the difference between the classical method and the proposed one. It is difficult to establish how much of the difference in the results between the proposed method and the classic one is due to the population bias (difference in scale) and how much is due to the quality of the method itself. Perhaps, replicating the research with different datasets could give enough information to understand how much the population scale decreased the quality of the classic method.

The second and equally important limitation of this research is the role of the pseudo-random generator. This is a potential source of error because, as shown by Van Niel and Laffan (2003), under certain conditions the pseudo random software generators can have cyclic behaviors. However, in this study the random generator used for both experiments with Monte Carlo was the same but we did not test the limits of the pseudo random generator. It is a minor flaw since using the same pseudo random generator the eventual error is the same for both the Monte Carlo.

Acknowledgements

Thanks to M. Enthoven and A. Brouwer for providing the necessary data for this research.

References

Aldstadt, J., 2010. Spatial Clustering. In: M.M. Fischer, A. Getis. (eds) *Handbook of applied spatial analysis* (pp. 255-278). Springer, Berlin, Heidelberg

Anselin, L., Cohen, J., Cook, D., Gorr, W. and Tita, G., 2000. Spatial analyses of crime. *Criminal justice*, *4*(2), pp.213-262.

Auchincloss, A.H., Gebreab, S.Y., Mair, C. and Diez Roux, A.V., 2012. A review of spatial methods in epidemiology, 2000–2010. *Annual review of public health*, *33*, pp.107-122.

Baddeley, A., Rubak, E., Turner, R., 2015. *Spatial point patterns: Methodology and applications with R*. CRC press, London.

Bonner, M.R., Han, D., Nie, J., Rogerson, P., Vena, J.E. and Freudenheim, J.L., 2003. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*, *14*(4), pp.408-412.

Boscoe, F. P. (2008). The science and art of geocoding: Tips for improving match rates and handling unmatched cases in analysis. In: G. Rushton, M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West, & D. L. Zimmerman (eds), *Geocoding health data: The use of geographic codes in cancer prevention and control, research and practice* (pp. 95–110). Boca Raton, FL: CRC Press.

Brix, A. and Kendall, W.S., 2002. Simulation of cluster point processes without edge effects. *Advances in Applied Probability*, *34*(2), pp.267-280.

Brown, J.D. and Heuvelink, G.B., 2008. On the identification of uncertainties in spatial data and their quantification with probability distribution functions. In: Wilson J.P. and Fotheringham A.S. (eds). *The handbook of geographic information science.* Blackwell Publishing, pp.94-107.

Clark, P.J. and Evans, F.C., 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, *35*(4), pp.445-453.

Cressie, N., 1991. *Statistics for spatial data*. John Wiley & Sons.

Daley, D.J. and Vere-Jones, D., 2003. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer New York

DeLuca, P.F. and Kanaroglou, P.S., 2008. Effects of alternative point pattern geocoding procedures on first and second order statistical measures. *Journal of Spatial Science,* 53(1), pp.131-141.

Diggle, P.J., 2013. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.

Duranton, G. and Overman, H.G., 2005. Testing for localization using micro-geographic data. *The Review of Economic Studies*, *72*(4), pp.1077-1106.

Hope, A.C., 1968. A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(3), pp.582-598.

Enthoven, M.P., Brouwer, A.E., 2019. Investigating spatial concentration of sustainable restaurants: It is all about good food! *The Annals of Regional Science* 1–20.

Gatrell, A.C., Bailey, T.C., Diggle, P.J., Rowlingson, B.S., 1996. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers* 256–274.

Getis, A., 2010a. Spatial autocorrelation. In: M.M. Fischer, A. Getis. (eds) *Handbook of applied spatial analysis* (pp. 255-278). Springer, Berlin, Heidelberg.

Getis, A., 2010b. Second-order analysis of point patterns: the case of Chicago as a multi-center urban region. In: Anselin L., Rey S.J., (eds) *Perspectives on Spatial Data Analysis* (pp. 83-92). Springer, Berlin, Heidelberg.

Goodchild, M. F. 1998. Uncertainty: The achilles heel of GIS? *Geo Info Systems* 8(11): 50–2.

Goodchild, M.F., 2010. Whose hand on the tiller? Revisiting "Spatial Statistical Analysis and GIS". In: *Perspectives on Spatial Data Analysis* (pp. 49-59). Springer, Berlin, Heidelberg.

Goldberg, D.W. and Cockburn, M.G., 2012. The effect of administrative boundaries and geocoding error on cancer rates in California. *Spatial and spatio-temporal epidemiology*, *3*(1), pp.39-54.

Goldberg, D.W., Wilson, J.P. and Knoblock, C.A., 2007. From text to geographic coordinates: the current state of geocoding. *URISA journal*, *19*(1), pp.33-46.

Hart, T.C. and Zandbergen, P.A., 2013. Reference data and geocoding quality: Examining completeness and positional accuracy of street geocoded crime incidents. *Policing: An International Journal of Police Strategies & Management*.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons.

Ingram, D.R., 1978. An evaluation of procedures utilised in nearest-neighbour analysis. *Geografiska Annaler: Series B, Human Geography*, *60*(1), pp.65-70.

Kirby, R.S., Delmelle, E. and Eberth, J.M., 2017. Advances in spatial epidemiology and geographic information systems. *Annals of epidemiology*, *27*(1), pp.1-9.

Kulldorff, M., 1998. Statistical methods for spatial epidemiology: tests for randomness. In: Gatrell A and Löytönen M (eds) *GIS and Health*. London, Taylor and Francis: 49–62

Lee, J. and Wong, D.W., 2001. *Statistical analysis with ArcView GIS*. John Wiley & Sons.

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

Owusu, C., Lan, Y., Zheng, M., Tang, W. and Delmelle, E., 2017. Geocoding fundamentals and associated challenges. pp.41-62. In: Karimi, H.A. and Karimi, B. (eds). *Geospatial Data Science Techniques and Applications*. CRC Press.

PDOK, 2021. PDOK, *Hét platform voor hoogwaardige geodata*. url: https://www.pdok.nl/

R Core Team, 2021. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ratcliffe, J.H., 2004. Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*, *18*(1), pp.61-72.

Ripley, B.D., 1977. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(2), pp.172-192.

Sahar, L., Foster, S.L., Sherman, R.L., Henry, K.A., Goldberg, D.W., Stinchcomb, D.G. and Bauer, J.E., 2019. GIScience and cancer: State of the art and trends for cancer surveillance and epidemiology. *Cancer*, *125*(15), pp.2544-2560.

Scott, L.M. and Janikas, M.V., 2010. Spatial statistics in ArcGIS. In: M.M. Fischer, A. Getis. (eds) *Handbook of applied spatial analysis* (pp. 27-41). Springer, Berlin, Heidelberg.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), pp.234-240.

Van Niel, K. and Laffan, S.W., 2003. Gambling with randomness: the use of pseudo-random number generators in GIS. *International Journal of Geographical Information Science*, *17*(1), pp.49-68.

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*(1), pp.1-12.

WorldPop, The, 2021. *The World Pop project.* url: https://www.worldpop.org/

Yamada, I. and Rogerson, P.A., 2003. An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical analysis*, *35*(2), pp.97-109.

Zandbergen, P.A., 2008. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems,* 32(3), pp.214-232.

Zandbergen, P.A., 2009. Geocoding quality and implications for spatial analysis. *Geography Compass*, *3*(2), pp.647-680.

Zhang, J. and Goodchild, M.F., 2002. *Uncertainty in geographical information*. CRC press.

Züfle, A., 2021. Uncertain spatial data management: An overview. In: Werner, M. and Chiang, Y.Y. (eds). *Handbook of Big Geospatial Data.* Springer Nature.