Reflection on master thesis: “Empirical approach to the problem of point pattern analysis in case of geocoded data”

Alessandro Fois

List of contents

The decision process

This document contains some reflections and discussion related to the thesis entitled *Empirical approach to the problem of point pattern analysis in case of geocoded data*. It is separated from the main text because it has been chosen to follow the template that leads to the production of a draft scientific article to be submitted. Therefore, the main document is the paper draft, that contains all the information usually essential in a scientific article, the theoretical framework, the methodology, the results and the conclusions. While in this document, the following paragraphs explain the reasons that led me to deal with the issue of geocoded point distributions, the theories on which I based my arguments and the theories I discarded. In addition, there is a conceptual model that should help the reader understand the structure of the reasoning that governed the research process and the draft paper writing. Finally, there are some reflections on the difficulties or uncertainties that emerged during the work, on the role of supervisors in advising me on decisions.
The usual praxis requires a section dedicated to any ethical issues emerging or related to data processing such as anonymization. Since the data used was not affected by any restrictions, I only spent a few lines explaining in detail the reasons for the absence of the usual paragraph.
The document closes with an attachment of the logbook listing all the activities carried out, arranged chronologically. However, it does not include the actual errors that would not give any information except justify the time spent. Though, it contains the practices discarded because they were considered disadvantageous, with this a reader should be able to replicate the study easily or simply examine the code.

General discussion about the topic

The topic discussed in this thesis is relatively original. Relatively, because it does not propose something new *per se*, but questions the existing methodology and suggests a different approach, adapting the existing methods, to tackle the highlighted problem. As explained in the draft paper, it is argued that the geocoding process limits the sample space and this has to be accounted for by the further analysis methods applied. Consistently with the hypothesis, is proposed an approach to calculating the expected value of the average distance in conditions of randomness when the observed pattern of points has been geocoded.
The idea developed in a specific context that influenced subsequent choices. During my internship at the University Medical Centre in Groningen I participated in a geo-epidemiological research work, which, as in this study, used various techniques of spatial analysis. That research is not yet published and therefore cannot be referred to in this text. However, it contained some of the difficulties common to many

epidemiological analysis located in developing countries or complex contexts. Some of these were, the geocoding information not always well recorded, the lack of high quality, up-to-date and equally detailed geographic information throughout the territory, inadequate information on population density, and ultimately, the observed pathology transmitted by the mosquito-vector Aedes Aegypti, was geographically represented by the residence of the infected person. These, and in particular the latter, are very frequent limitations in epidemiology, largely discussed in the literature, as they have potentially serious consequences on the reliability of the results and therefore of the conclusions (e.g. Gatrell et al., 1996). However, a further peculiarity was added to this context which led me to reflect more carefully on the concept of randomness and test applications. The study used data from the first spread of the disease in that area, which being an island of limited size could be considered completely new to the virus and thus, lacking in 'immune experiences'. This, unless other indicators are observed, makes all the population equally exposed, meaning that the first wave diffusion would most likely have followed a completely random pattern, thus events independent each other (Diggle, 2013, p99) or at least a random pattern approximated with a Markov chain (e.g. Yaesoubi and Cohen, 2011; Dehghan Shabani and Shahnazi, 2020).

Nonetheless, the results continued to show areas of hotspots and the pattern showed clear non random space-time progressions. Therefore, in an attempt to understand if that was an error or if simply that was the pattern dynamic and the population was not equally exposed, I noticed that the distance-based tests considered the space uniform, albeit within certain limitations, while the points could only be placed in specific places. In other words, while the geocoding process was used to transform the event list into a map, to test the null hypothesis was generated a number of random patterns that simulate a homogeneous Poisson process, following the modellization of the expected behaviour of the virus, independently from the geocodification process. Even considering the population density, which largely reflects the density of houses, a disparity remains. Because it was considered all the available surface as a potential place where an event could occur. This fact, at least intuitively, unbalanced the outcome of the test strongly in favor of the rejection of hypotheses.

In fact, if a geocoded address is represented by a point ideally located, for example, in correspondence with the house number, or in the center of the building, that single point represents the entire surface of the property. However, when a random distribution of points is generated without considering this fact, not only can a property contain more than one address, but the possibilities for distance fluctuations are much greater, infinite in theory, than those of geocoded points, which although numerous are countable. It could be objected that it is in the nature of the random process to find an equilibrium that compensates for the disproportion of the locations, indeed, if a process is random so a stationary Poisson point process, every point location of the

surface sample is equiprobable as far the mean and variance are equal to any other surface sample (Daley and Vere-Jones, 2003, p19-26). Therefore, over the same surface, a possible concentration in correspondence of a single property will be balanced by an opposite disparity somewhere else.

However, from this it can be deduced that the surface has a great relevance and unrealistic or impossible areas, compared to the observed event, should be excluded. This series of considerations reinforced the idea of a test to clarify any doubt on possible invalidation of the methodology due to geocoding. Therefore, it was first necessary to verify that under conditions similar to those of the observed pattern, the test was really able to discern a true random point pattern. To do this, it was necessary to 'geocode a random pattern', in other words, randomly select a sample of addresses from the geocoding database. If the database is not accessible the test cannot be done, so in order to perform the test it was necessary to move to a geographical area where the data were available.

Since the geographical area was no longer that from the epidemiological analysis in which the idea emerged, a case study was needed in a place with the necessary data, such as the Netherlands. Furthermore, a study case was needed that made use of one methodology of analysis of points based on the distance. This turned out to be a more difficult task than expected, because the research engines can only search between the keywords, and this was not enough to retrieve articles where geocoding in a place with high quality data and spatial analysis with distance based methods coexist. Eventually, knowing one of the authors it was possible to obtain the data of the paper by Enthoven and Brouwer (2019) which had the required characteristics. Furthermore, it uses the Clark-Evans test (1954) which has the advantage of being easy to implement. Although other tests, such as Knox space time (1964) or Ripley's K function (1979) are more used, they follow a not as adaptable approach to the space concept of this hypothesis compared to the Clark-Evans test. Overall, this is not a significant limitation given that if the hypothesis had not been confirmed, even with one simple test, then it would not have been of great interest to continue working to implement a variant in the more complex ones.

Finally, it must be said that during the writing of the draft of the paper, it started to strengthen the thought that a purely theoretical approach could suffice to demonstrate the validity of the hypotheses. However, in order to demonstrate a mathematical methodology, or perhaps confutate it, are required advanced mathematical and statistical skills. In addition, an experiment still remains a valid and useful effort, even more so after the first results have been obtained, which have given a possible measure of the degree of approximation between the two methods (not easy to obtain with a theoretical approach). Indeed, with reference to the needs of the epidemiological study I mentioned above, it would not be possible to apply the proposed method because of lack of data. However, knowing within which range of values the ordinary methods are

reliable would be a great added value. Clearly, a more detailed knowledge about the difference between the classic approach and a proposed one can be seen as a benchmark to assess under which condition the classic approach represents a good balance between data availability and quality of the measurement. At the moment is not realistic think each research that make use of geocoding has access to the spatial database where is stored the information. Until that moment, even if the proposed method is more effective than the classic one, it would be a great achievement if it was possible to collect enough experiments, with more case of studies from different contexts, to see if the error stay inside a range so that it could be used to adjust the classic method.

Other theories or methods

This thesis work investigates how the geocoding procedure affects point pattern analysis methods. In doing this it focused on a specific distance based method, chosen also for its simplicity, and in particular on the calculation of the expected value of a homogeneous poisson process, synonymous with randomness (Ripley, 1977; Cressie, 1991, p663; Illian et al., 2008, p57-58). Therefore, it is a low-level approach, just above statistical theory. This is confirmed by the wide range of disciplines that make use of both the geocoding procedure and the distance-based spatial analysis methods. As a consequence, the choices fell not on which family of methods to choose (rejecting others) to explain an observed phenomenon, but on how to maintain a broad and inclusive approach even making use of specific examples such as the case study.
Following this purpose, some aspects of basic spatial statistical theory, common to all disciplines that make use of geocoding and spatial analysis, were first explicitated. This resulted in choosing the texts of Cressie (1991) to which the terminology used in this work refers, Diggle (2013), Illian et al. (2008) and to a lesser extent (Daley and Vere-Jones, 2003), do not appear the remarkable book of Haining (2003) even though I found it useful as further confirmation of the theories, also from a chronological point of view. Ultimately, the main operative and recent text was the book of Baddely et al., (2015) that I used as a base for deriving my variation from the classic methodology. Some of these theories refer to the field of environmental geography other than spatial epidemiology, while Cressie (1991) remains the most basic and thus inclusive for all the applications. All this books and theories are among the most often cited, frequently cross-referenced; this could highlight the missing of *Spatial Econometrics, Methods and Models* by Anselin (1988), an important book but characterized by a strictly econometric approach, as the author himself states (Anselin, 2010). The combination of all of these sources give a broader view useful to define an appropriate theoretical framework. This was not easy, precisely due to the widespread diffusion of geocoding in different fields

which, from time to time, adapt the methodologies to their context. This could create conflicts with the assumptions made. For example, in epidemiology a random process could be influenced by the previous event and therefore violate the independence assumption of the Poisson point process, but the randomness' still be limited to a geographical space surrounding the original event and can be approximated by other models. One possibility would therefore have been to concentrate in detail on a specific case study, according to the practice of its field of application, but this would have restricted and complicated the applicability of the conclusions to other fields that make use of geocoding. Here the main elements are geocoding, distance-based methods and surface significance. According to the application field, therefore, the value of the conclusions could be different, although the presence of a relationship between these three elements remains clear.

In criminology, for example, according to Anselin (2000) it is known that there is a great discrepancy between the surface where events can occur, usually the city, and the space where the greatest concentrations of events are observed. This concentration is referred to as the crime place theory (Eck and Weisburd, 1995), which over time has consolidated its validity so much that Weisburd (2015) refers to it as the 'first law of the criminology of place'. Clearly any geographer is led to a parallel with Tobler's first law of geography (1970). Both have in common that they postulate a hierarchy and spatial significance that if accepted 'remove' the need to verify the potential presence of randomness. According to some authors, tests of complete spatial randomness are in fact meaningless, unless having no clue on what process is in progress (Aldstadt, 2010). Goodchild expressed very clearly how acceptance of Tobler's law implies spatial autocorrelation (2010). On the other hand, Miller (2004) observes that although Tobler's law remains important, the geographic conditions and concepts of distance have changed over the years, making measurement of spatial autocorrelation and spatial heterogeneity more important than before.

The existing models to measure the characteristics of spatial variables and their correlations are numerous, because there are numerous shades of reality; as Miller (2004) points out, geographical locations have uniqueness that become increasingly important together with the increasing complexity of society. According to this point of view, it is very difficult to accept a postulate that presupposes a sort of a priori spatial autocorrelation. And consequently it is not possible to establish without a measure of some kind, that an observed event could be a random realization. Consistently, most of the models are tested against the hypothesis of randomness which is frequently based on the homogeneous process of Poisson (Ripley, 1977; Illian et al., 2008).

It is therefore clear that simple variations in the theoretical assumptions used to justify the validity of the geocoding process are enough to change the meaning of the observed conditions. As also reported in the main document, the limitations introduced by the geocoding process are widely discussed in the literature, both in

epidemiology and in criminology (e.g. Zhan et al., 2006; Oliver et al. 2005; Summerton, 2015; Hart and Zandbergen, 2013). However, if in light of all the limitations of the geocoding procedure, the point pattern is believed to be representative of the observed phenomenon, then the question raised in this paper remains almost unanswered by the current literature. Almost, because a family of theories that was not included in the main document is the one that applies the point pattern process to the linear phenomena (e.g. Baddeley, et al., 2021; Moradi and Mateu, 2019; Ang et al., 2012; Okabe and Satoh, 2009). The reason why it was excluded is the 'need' to tackle the problem from a two-dimensional point of view, which best approximates many of the conditions in epidemiology and other disciplines. Moreover, since a linear network is not a homogeneous environment, given the non-uniform distribution of the lines, understanding the relationship between the points is more complex compared to the Euclidean space case (Badeley et al., 2015, p711). Similarly to what has been achieved in this work, also the research about point patterns on linear networks worked in the direction of transforming the existing methodologies adapting them in accordance with the space. However, some phenomena are universally recognized as being part of the networks, such as road accidents, street crime or retail assessment (e.g. Araldi and Fusco, 2019; Khalid et al., 2017; Fan et al., 2018), and the problem does not lie (primarily) in the geocoding if used, but in the conceptual link between the observed events. Okabe et al., (1995) provided a clear example case of the reseller in the city linked to their customers through the road network not the Euclidean distance that would mean flying.

On the contrary, in the problem addressed in this study it is necessary to first demonstrate that the geocoding procedure alters the relationships between distances, and that it is an acceptable limitation only under certain conditions. Researchers who have tackled the problem of networks have also chosen to adapt a distance-based function, among the first in this field it should be mentioned Okabe and Yamada (2001), who adapted the K-function to the networks, and Okabe, Yomono and Kitamura (1995) which in a much more advanced way, adapted the Clark-Evans test to the networks.

Although research on point patterns in linear networks has developed from tackling problems in a similar way to what has been done in this work, as the distinction between continuous and discrete space (Okabe et al., 1995) and adapting the distance based functions, since they are the most used (Okabe et al., 1995; Okabe and Yamada 2001; Ang et al., 2012; Moradi and Mateu, 2019), it is believed that placing them in the theoretical framework could have created confusion with respect to the concept of 'transformation of space' due to geocoding. It seemed that starting from an established and shared theoretical framework such as that of Cressie, and deriving both the variant of the networks (and possibly the geocoding on lines) and that of the surfaces (given by the geocoding), could add confusion to a sufficiently articulated topic.

Finally, the theories of spatial analysis that make use of spatial units were excluded from the discussion, in particular Moran's I and Getis G*, which were used by the Enthoven and Brouwer paper used as a case study. In this regard, the choice was determined by the need to exclude the results applied to methods not directly based on points. As is known, a pattern of points can be grouped into spatial units, which are polygons that usually have some meaning that justifies the grouping of the observed variable. The characteristics of the spatial units have an effect on the autocorrelation analysis (Anselin and Getis, 2010) and therefore should be justified with a theoretical framework as well. Therefore, even in this case the inclusion of these theories would have firstly increased the complexity, secondly it would have extended the discussion beyond the limits allowed by scientific publications and finally, it would have pushed forward the research from the objective of demonstrating the bias introduced by geocoding. Moreover, inserting these parts in a coherent way would have caused the necessity to replicate the research of Enthoven and Brouwer, following its theoretical context and any limits, risking to go off topic by transforming a research that tries to answer a research question into a critique of another work. Contrairly, the purpose was to show the applicability of the proposed method to a real case.
Finally, in this regard it should be mentioned that with regret it was not possible using one of the classic case studies, which are usually recommended for this type of research. This was due to the difficulties in finding a geocoding database consistent with the time and location of the potential case of study. For instance, the famous paper by Getis and Ord (1992) introducing the G statistics, made use (also) of the data presented by Cressie and Chan (1989) to take advantage of the possibility of comparison between the methodologies, in fact, these common and classic datasets are considered as benchmarks.

Excluded results

The geocoding process is often used even if the pattern is not expected to be analyzed with point pattern methods. The case study is a particularly advantageous example because it combines the use of point pattern analysis methods with different approaches, based on spatial units. Therefore, when the data was already loaded on the computer ready to be analyzed, it was very simple to extend the analysis to other methods. The idea was to understand if the bias found with respect to the Clark-Evans method was only a defect of this methodology or if the other methodologies also present anomalies. The same random point patterns generated to test the proposed method and the classical method were loaded on the spatial units defined for the case study. In this way it was possible to compare how the combination of spatial units and spatial autocorrelation tests such as Moran's I and Getis G* behaved with respect to the values obtained in the case study. In all the tests carried out, the autocorrelation values

were much greater than those of the case study. This was actually a predictable result, because the practice of creating spatial units by trying to find the optimal distances or dimensions of the spatial units, so that the spatial autocorrelation values are maximal, simply picks up variability from space. Therefore the result was a false positive. A solution to this problem could be the use of a mathematical function that defines the shape of a polygon based on the value of a variable. As a further check, the data has also been loaded on the spatial units that represent the administrative boundaries. Also in this case the result was in line with expectations. The random pattern (values once loaded in the spatial units) were strongly autocorrelated. The population is probably correlated with the municipalities, so a random sample of the population loaded into the administrative boundaries does nothing but replicate the density ratios between municipalities.

Inserting these results in the main document, would have changed the objective of the research. From identifying and possibly measuring a bias due to geocoding to a discussion on bad practices in the use of autocorrelation tests. Furthermore, these are distinct topics, the presence of a bias due to geocoding in distance-based methods is not immediately evident, while it immediately raises some doubts that generic spatial units such as administrative boundaries can introduce a bias. Because these are polygons without a specific relationship with the observed variable and which, however, have been defined in another context following some logic. They could follow, in the simplest of hypotheses, the geography of the place.

However, I did not perform the same experiment described above with the classic Monte Carlo method. In this case the result is not easily predictable. I expect that a random pattern generated by choosing random points across the surface can be more easily recognized as random than a geocoded one, especially when loaded into regular spatial units such as squares. I am not so sure that the administrative boundaries are capable of such a distinction, probably after many attempts, loading many different random patterns, one could understand the probability of success. This is, perhaps, because the size, shape, and configuration of the spatial units affect the result of the autocorrelation test (Anselin and Getis, 2010, p40).

The journal choice

This thesis work involves the writing of two documents, a draft of a scientific article, and this document which contains the reflections that usually are not part of a scientific article. Subsequently the draft must be sent to a newspaper to request publication. As is well known, there are numerous scientific journals, with different themes and different impact factors. This establishes a hierarchy in journals, researchers try to publish in high impact journals, and journals try to select authors who can offer the most professionalism and share. Since the submission and

evaluation procedure can become a source of frustration, it is commonly recommended to write a draft paper following the guidelines and characteristics of the journal that is deemed most appropriate for your work.

The topic addressed is a hybrid between statistics and geography, sometimes addressed as spatial statistics. However, there are mainly elements of what in the field of GIScience could be defined computational geography (I believe that if we call computational physics the set of computer procedures that allow the calculation of physical-mathematical models then the geographical analogue should be called computational geography as a specialization of GIScience) than purely statistical reflections. Therefore, if a purely statistical approach had been chosen for the solution of the raised problem, then journals such as *Journal of Applied Probability* or *Journal of the Royal Statistical Society: Series B* might have been appropriate. However, the issue was dealt with mainly geographically, so the choice should fall on newspapers that have GIScience as their main theme.

There are numerous newspapers of this type, and among these, I believe the most prestigious and attractive is the *International Journal of Geographical Information Science.* Without forcing the presence of citations to articles from this journal, there are two references from it in the draft article attached to this document. They might seem few, but there are no journals with more than two references in the draft and this could leave room for some shy expectations. However, realistically speaking, being able to publish in this journal, with a really high impact factor and where some of the most important authors in the field of GISciences have published, and do publish, is not very likely. A reasonable strategy is to choose a journal with a smaller but equally prestigious impact factor, such as *Journal of Geographical Systems*; obtaining a publication in this journal would be equally important. Furthermore, it would have the advantage of obtaining high-level feedback. If at the end of the peer review procedure the outcome should still be a refusal, the value of the suggestions obtained would still remain, and the draft article could still be submitted to a prestigious journal, other than this time could take advantage of the advice from the first journal.

Other reflections

During the processing of this thesis, numerous doubts arose on the validity of the proposed methodology, on the research question and on the approach followed.

On the one hand there was the evidence of the results that was reassuring. I replicated a case study and at the same time I showed that the same methodologies were not able to identify what they are supposed to see, randomness. On the other hand, these results seemed so obvious that I wondered if it is simply a wrong and widespread practice in research, despite the peer review procedure. For example, choosing spatial units

without a specific criterion just because they are the only ones available is equivalent to manipulating data, I believe. This is why all the random patterns generated were autocorrelated.

I began to believe that my work was wrong, or at least that it could be traced back to a simple theoretical demonstration, even based on a literature review. Of course the result of a test has a very strong value for proving a hypothesis, but it would remain a mere exercise compared to possible existing theories. In fact, even from the theories gleaned from the literature review, it emerged that many of the frequent practices are inappropriate. For example, it is inappropriate to apply a model that assumes the presence of spatial homogeneity when it is known a priori that the observed pattern cannot be homogeneous.

These reflections forced me to question my knowledge on the subject. I felt the need to explore the subject even more and I believe it requires advanced mathematics skills that are not part of a geographer's cultural background. I wondered for a long time whether the research question was as appropriate as it seemed when it was conceived. However, at the end of the writing process, the research question seems to me a rhetorical question that does not deserve an answer. Perhaps, a solution to these doubts, in anticipation of a future publication, would be to divide the writing into two distinct parts. One theoretical article, which discusses the structure of the spatial database with respect to the conditions of independence of the points and equiprobability given by the homogeneous distribution of poisson. And a second article in which the previously discussed theory is applied, and the method derived from the new theoretical approach is tested with a case study.

Why no ethics here

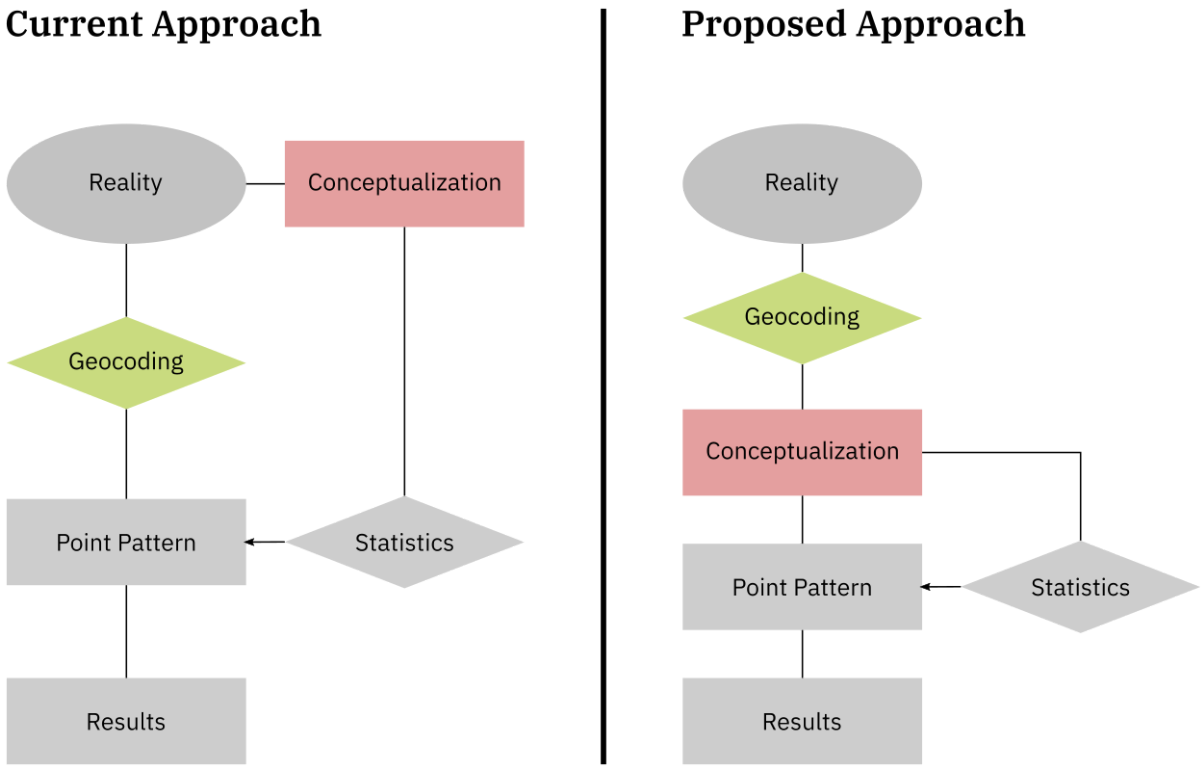
Ethics was not an issue in this research. The research does not involve participants, it does not discuss sensitive topics (Lee and Renzetti, 1993), it does not use personal data, neither the case of study does, since the restaurant locations are in public places. Lastly, it did not benefit from any funding nor did the author have connections or can take advantage with/from any of the arguments discussed.

However, ethics and privacy are not the same thing, especially when we distinguish ethics from morality (see Shaw, 2015; Moore, 2008). Unless we do not care to fall into the contradiction in terms of 'ethics and science' as argued by Weber (1949). In respect to the current discussions on scientific research and ethics it would seem a reference to an ethics of privacy, but it would be more appropriate to speak of the morality of privacy. Although morality of privacy could be also inappropriate, since from a research operational point of view, thus, the act of observing behavior or personal data, has more easily to do with privacy as a tool that has its own intrinsic value (Moor, 1991). However, if the relation between ethics and privacy lies in 'ethics in the formation of

trust', and privacy is the capability of every subject to control its own personal data then the issue is to act in a fair and predictable manner (Goldberg et al., 2001).

Conceptual model

This research is based on comparison. It is a question of comparing the results of the proposed methodology with the ordinary ones. As evidenced by the research question, *Does the geocoding process violate the assumptions of the (in) homogeneous poisson process?* The object is whether geocoding introduces a bias with respect to the conceptualizations that determine the abstraction of reality and its description through a statistical model. Simply put, does geocoding make space discrete with fixed and countable locations or not? Hence, it's just a way to assign coordinates in continuous space?
The conceptual model is shown in Figure 1, the comparison between the results generated by the ordinary approach and the one proposed is symbolized by the comparison between the two logical models. The difference is symbolized by the arrangement of the geocoding procedure, independent from the conceptualization in the ordinary model, and preceding the conceptualization in the proposed model. Of course, as also repeated in this document, the limitations introduced by geocoding are widely discussed in the literature, so it could be argued that the conceptualization is subsequent to geocoding in both models. However, no literature has been found discussing the limitations of geocoding with respect to the question raised in this work. Thus, we can consider the schema representative of the current context with respect to the research question.
The subsequent conceptualization of geocoding indicates that reality, represented by a point pattern, is also abstracted and modeled through the geocoding process, to which it ascribes the side effect of altering the initial model from continuous in space to discrete. Consequently, this new condition is included in the statistical model where the change is represented by the chain of definition in cascade, thus the statistical model follows the construction of the pattern and includes its characteristics.



The research process

Reflection on the research process and role of supervisors

The development of this research was a very interesting and formative experience. Having chosen the Research Master it was not a new procedure but the combination of many procedures learned during the two years of study. Looking back at the whole process, I have a pleasant memory, it was the first assignment that could be carried out without other tasks to be performed. This meant that time could be organized as a function of research. I find it extremely unproductive to overlap different and complex works, even if, after a long period in which we always work on the same theme, a pause is necessary in order not to freeze on the same positions. Despite this, the research did not start when all the courses were finished, indeed its very beginning was about a year ago during the internship. At the time I didn't know it was going to be a thesis topic, it was just a problem to be solved. Gradually and slowly the concepts took shape, also thanks to the courses of the master. Two of these in particular played a decisive role, the 'Research proposal and proposal writing' course and that of 'Entrepreneurship and regional development'. In the first one, I had the opportunity to reflect on the concept of randomness in research, an extremely complex and fascinating concept. I have been able to observe that some procedures that make use of the concept of randomness do

not contrast two opposites, order and chaos but observe the predictability of an event. Furthermore, this vision of randomness linked to the predictability of an event is not at all obvious nor common to all disciplines. After all, this research speaks more of randomness than geocoding, although it is not expressed in these terms, the research question asks precisely whether geocoding interferes with randomness. In the second course, 'Entrepreneurship and regional development'. I was able to learn about the case study on sustainable restaurants by Enthoven and Brouwer, in that article observing the chain of analysis and procedures, generating the need to argue and contextualize the hypothesis of this research even outside of epidemiology.

Clearly, my supervisors were crucial. Daniella Vos has the gift of being able to motivate you even when you just want to give up. Suddenly everything seems simple to you, then when you realize that it is not at all like this, the worst is over.

I am grateful to my second supervisor Michiel Daams, who is always very present. I appreciated and benefited from his precise and timely advice. In addition, the entire period of thesis work took place during the pandemic, so it was not possible to meet in person and ask for advice or exchange opinions is less spontaneous. You can't just knock on an office door or chat during a coffee break, every request must be written or planned and this gives a perception of bothering. Personally I think that, writing an email or planning a meeting, is like overloading a person who is already busy enough with further work, not only towards my supervisors. Nonetheless, we managed to communicate well enough, I think. Certainly I really appreciated how both my supervisors managed a moment of despair mainly due to the condition of isolation combined with some failures; in a different condition I could have given up. Finally, my supervisors were decisive in helping me decide the topics to cover in my presentation. Normally I do not have any problems in presentations, moreover it is not a new experience for me. However, in this circumstance there were a combination of factors that all together made the presentation a new experience as well as difficult. The duration of 7 minutes, the videoconference with a connection that is not really performing, in combination with the paradigm shift with respect to time management. In my experience it is essential to contact the listeners, to understand if they are following you and to be flexible in indulging on the topics that interest most and to overlook those that arouse any interest, even at the cost of taking more time than expected. listeners to decide. All this was not possible, because here the paradigm provides for meticulous planning of the topics in order to stay on time regardless of the reactions of the public. I believe, especially after the listeners' feedback, that the presentation went well, and this would not have been possible without the help of my supervisors and Daniella in particular.

Finally, I really appreciated the flexibility of both of them with respect to the structure of the research project. I approached them with a draft research proposal that certainly needed further study, but on the other hand I had a well-structured methodology and

the first timid experiments to test my hypothesis. In this context they were able to reorganize the procedure so as not to have to interrupt what was done, allowing me to obtain results and then proceed with the writing.

Appendix: The logbook

- *See the attached document* -

References

Ang, Q.W., Baddeley, A. and Nair, G., 2012. Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics*, 39(4), pp.591-61

Anselin, L., 1988. *Spatial Econometrics: Methods and Models* (Vol. 4). Springer Science & Business Media.

Anselin, L., 2010. Thirty years of spatial econometrics. *Papers in regional science*, *89*(1), pp.3-25.

Anselin, L. and Getis, A., 2010. Spatial statistical analysis and geographic information systems. In: Anselin, L. and Rey, S.J., (Eds). *Perspectives on Spatial Data Analysis* (pp. 35-47). Springer, Berlin, Heidelberg.

Araldi, A. and Fusco, G., 2019. Retail Fabric Assessment: Describing retail patterns within urban space. *Cities*, 85, pp.51-62.

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. and Davies, T.M., 2021. Analysing point patterns on networks—A review. *Spatial Statistics*, 42, p.100435.

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

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.

Cressie, N. and Chan, N.H., 1989. Spatial modeling of regional variables. *Journal of the American Statistical Association*, *84*(406), pp.393-401.

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.

Dehghan Shabani, Z. and Shahnazi, R., 2020. Spatial distribution dynamics and prediction of COVID-19 in Asian countries: spatial Markov chain approach. *Regional Science Policy & Practice*, *12*(6), pp.1005-1025.

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

Eck, J.E., and Weisburd, D. 1995. Crime places in crime theory. In: Eck, J.E. and D. Weisburd (eds). *Crime prevention studies.* Monsey, New York: Criminal Justice Press

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.

Fan, Y., Zhu, X., She, B., Guo, W. and Guo, T., 2018. Network-constrained spatio-temporal clustering analysis of traffic collisions in Jianghan District of Wuhan, China. *PLoS one*, 13(4), p.e0195093.

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., Ord, K.(1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24:189-206. Reprinted with permission of Blackwell Publishing, Oxford

Goldberg, I., Hill, A., and Shostack, A. 2001. Trust, ethics, and privacy. *Law Review* Boston University, 81(2), 407-422.

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.

Haining, R., 2003. *Spatial data analysis: theory and practice*. Cambridge university press.

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.

Khalid, S., Shoaib, F., Qian, T., Rui, Y., Bari, A.I., Sajjad, M., Shakeel, M. and Wang, J., 2018. Network constrained spatio-temporal hotspot mapping of crimes in Faisalabad. *Applied Spatial Analysis and Policy*, 11(3), pp.599-622.

Knox G., 1964. The detection of space-time interactions. *Applied Statistics*, 13, 25-29.

Lee, R.M. and Renzetti, C.M., 1993. Researching sensitive topics.

Moor, JH , 1991. The ethics of privacy protection. *Library trends*, University of Illinois

Moore, A. 2008. Deining Privacy. *Journal Of Social Philosophy*, Vol. 39-3, 411–428

Moradi, M.M. and Mateu, J., 2019. First-and second-order characteristics of spatio-temporal point processes on linear networks. *Journal of Computational and Graphical Statistics*, 29(3), pp.432-443.

Okabe, A. and Satoh, T. 2009. Spatial analysis on a network. In A.S. Fotheringham and P.A. Rogers, (Eds). *The SAGE Handbook on Spatial Analysis*, chapter 23, pages 443–464. SAGE Publications, London.

Okabe, A. and Yamada, I., 2001. The K-function method on a network and its computational implementation. *Geographical analysis*, 33(3), pp.271-290.

Okabe, A., Yomono, H. and Kitamura, M., 1995. Statistical analysis of the distribution of points on a network. *Geographical Analysis*, 27(2), pp.152-175.

Oliver, M.N., Matthews, K.A., Siadaty, M., Hauck, F.R. and Pickle, L.W., 2005. Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics,* 4(1), pp.1-9.

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

Ripley, B. D. (1979). Tests of 'randomness' for spatial point patterns. *Journal of the Royal Statistical Society B*, 41, 368-374.

Shaw, R.M., 2015. Defining Ethics and Morality. In *Ethics, Moral Life and the Body* (pp. 11-36). Palgrave Macmillan, London.

Summerton, D., 2015. The Impact of Street Geocoding on Crime Hot Spot Prediction.

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

Weber, M. 1949. “Objectivity” in social science and social policy. In Weber, M., Shils, E. and Finch, H. A. (Eds). *The methodology of the social sciences*. New York: Free Press.

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

Yaesoubi, R. and Cohen, T., 2011. Generalized Markov models of infectious disease spread: A novel framework for developing dynamic health policies. *European Journal of Operational Research*, *215*(3), pp.679-687.

Zhan, F.B., Brender, J.D., Lima, I.D., Suarez, L. and Langlois, P.H., 2006. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of epidemiology*, 16(11), pp.842-849.