Geo-Media and Neighbourhood Effects' Studies: A New Frontier

Ciro Clemente De Falco

Department of Social Sciences, University of Naples "Federico II", Naples, Italy

Email address:

Ciroclemente.defalco@unina.it

Abstract: In the era of data revolution, new types of data and data sources allow researchers to find innovative ways to study society and its dynamics. The concept of neighbourhood effects (NE) was born within the sociological debate on the relationship between territory and social phenomena inaugurated by Durkheim and continued by other authors from many disciplines. NE is a particular concept that is borne more by empirical evidence than theory. This aspect is quite problematic because it is easy to find studies that investigate how spatial characteristics influence social phenomena; however, there is no agreement on the ways in which spatial influence manifests itself or on which spatial elements have a sociological bearing. Initially absent in Internet studies, NE have been progressively investigated owing to the geo-media. Crowdsourced Geographic Information has made it possible to jointly analyze two dimensions previously considered incompatible, such as the online and offline worlds. The purpose of this study is twofold. The first objective is to analyze how the discourse on NE is evolving in digital studies that use crowdsourced Geographic Information. The second objective is to identify the critical elements of this approach. In particular, we will try to give the reader the answer to the following questions: What kind of geo-media has been used? Which topics do NE research focus on? What are the hypothesized mechanisms that link space and social phenomena? What are the most frequently used approaches for this purpose? A systematic literature review was used to answer these questions.

Keywords: Neighbourhood Effects, Crowdsourced Geographic Information, Social Media, Ecological Analysis, Literature Review

1. Introduction

A neighbourhood is an area where people live and interact with each other and is characterized by social, ecological, cultural, and political factors. According to Sampson and colleagues [50] neighbourhoods are units nested within successively larger communities.

Since neighbourhoods are one of several central settings in everyday life, it is possible to assume that they can, directly and indirectly, influence the choices and life course of people who live there [51]. Exploring this assumption and understanding how the neighbourhood influences individuals' attitudes and behavior is the main objective of research on the so-called Neighbourhood Effects (NE).

The current NE approach is often associated with Wilson's [63] work entitled "'The Truly Disadvantaged", which deals with the effects of living in poor neighbourhoods [50, 61]. It is important to highlight that the relevance of spatial

characteristics on individuals had already been classified and/or investigated before Wilson's work. At the end of the 19th century the concepts of social morphology and density, developed by the French sociologist Émile Durkheim, embedded already the territory, its composition, and its effects on individuals [35]. In the 1920s, the American Chicago School focused on the city, its areas, and the social processes and mechanisms that ran through it [6]. Later, in the 1940s, the Swedish researcher Tingsten [56] published a work to analyze spatial effects on electoral behavior. In this last research thread, the interest in the territory's role was remarkably successful, as shown by Cox's [8] studies on NE in the 1970s and subsequent studies by Miller [43] and Johnston [28]. Today, NE is investigated in many fields, including medicine and epidemiology [5, 37].

In the literature, it is possible to find research on contextual,

spatial, compositional, and structural effects. Although they belong to different research traditions, these concepts are sometimes used interchangeably with the concept of NE. Semantically, the concept of NE could better represent the properties of the investigated territory. It is possible to discuss the ideal size of a neighbourhood to detect its effects [31, 53], but the reference to the object studied is quite clear. Do exist, instead, more 'contexts' and 'spaces' [54] in which the individual can be placed (work, family, friendship). While, with 'structure' in the social sciences it is understood a different object [10]. Eventually, the concept of composition, although used to identify a specific type of NE, does not refer to the territorial sphere. Moving beyond the semantic questions, the debate around NE focuses more on the appropriate methodologies and statistical techniques to detect them and on the type of mechanisms that can generate them, than on definitional issues. To detect neighbourhood effects, ecological analysis, experiments, and natural experiments can be used. However, it is widely recognized that experiments and natural experiments are hardly ever used [48], so do research on NE mainly means to analyze aggregate data, and indeed, a relevant part in the debate regarding the NE concerns the development of statistical techniques capable of detecting them correctly.

This development occurred hand in hand with the techniques that aimed to make inferences about individual behavior from aggregate data – such as the regression model proposed by Goodman [30] - without stumble into the methodological trap of the ecological fallacy identified by Robinson [26].

From Menzel's [41] response to Robinson's attack on the use of aggregate data - which he considered useless - it is possible, in fact, to draw indications on the correlation coefficient as a possible technique to use in order to hypothesize the existence of NE.

In addition, other than correlation, more sophisticated techniques have been proposed, even though they suffer from the problem regarding the separation of the analysis levels. In order to be able to isolate neighbourhood effects, there is the need for a technique that considers both the individual and aggregate levels at the same time [13]. The answer to this question was a multilevel analysis, that marked a turning point in the study of NE. Clearly, multilevel analysis remains a limited option to use because the analysis depends on the data availability.

Another relevant aspect for the analysis of NE is the study of mechanisms that relates neighbourhood to individual actions. An overview of the mechanisms was provided by Galster [16], who identified four families encompassing 15 types of NE. The first family is Social-Interactive Mechanisms, which refer to social processes that are endogenous to neighbourhoods. The second family is environmental mechanisms that refer to natural and human-made attributes of the local space that may directly affect residents' mental and/or physical health without affecting their behaviors. The third family "geographic mechanisms" refers to spaces' aspects that may affect residents' life courses. However, these aspects are the result of the neighbourhood's location in relation with larger-scale political and economic forces. Eventually, the last family regards actions enacted by whom typically not live in that given neighbourhood but have the control on important institutional resources and/or on points of interface between neighbourhood residents and vital markets. Identifying the mechanisms underlying statistical evidence is not straightforward, because in most cases it is necessary to involve qualitative investigations, that are more expensive than the quantitative ones for the research budget.

Despite the wide-ranging reflections that have developed around the NE and the empirical evidence gathered over time, there is still a debate regarding their existence. This is not surprising since the neighbourhood and more generally the territory - except for the classics mentioned above - have never played a relevant role in sociological elaboration. This absence regarding the space in the history of social theory and sociology itself, has led, in the twentieth century, some authors to talk about aspatial sociology [40].

However, the knowledge accumulated in this strand of studies, through the connection with the processes of datafication and the explosion of big data, can now create more value. Currently there is, in fact, an increasing availability of geo-localized data, that with the suitable theoretical premises, may bring out information on social phenomena. This growing amount of geo-localized data could open up new scenarios for NE. Since that much of geo-localized data come from social media, now seems possible to combine two worlds previously imagined completely separate: online and in-vivo.

This research aims to explore this possibility and its implications by reviewing the literature on geolocatized data.

In particular, the aim is to verify the existence of an NE approach using geo-data from social media and, if it is possible, its maturity, prospects and possible gaps. This could be useful for assessing the necessary steps for further analyses and for evaluating if geo-data analyses are data-driven or theory-driven. This aspect is not trivial because allows the comprehension on how the new information context influences the epistemological level of the research.

Scholars who are committed to ecological and NE analysis are aware of the problem of sources, their updating and consistency. Therefore, it is important investigate which territorial variables are used as well as the analytical approaches used (quantitative or qualitative or mixed methods) to assess the maturity of this new line of research.

RQ1 Is there a NE-related approach working on geo-social data coming from social media?

RQ2 What are the geo-social data sources used?

RQ3 What aspects of the territory are taken into consideration?

RQ4 What type of approach is used?

The article is structured in the following way. In the second paragraph, a review of the types of geographic data from the platforms and from the web will be made to identify the ones suitable for the research. The methodology used for the

literature review will be described in the third paragraph. In the fourth paragraph the principal themes of the literature on geo-social data will be identified. Then, in the fifth and sixth paragraphs will be investigated 1) the research that uses geo-data in compliant perspective with NE, 2) the sources and types of geo social data used in the literature and 3) the approaches used and the characteristics of the territory.

2. New Data for NE Studies

In recent years, there has been a sharp increase in the volume, variety, and velocity of data from different sources. The big data concept is typically used to describe this vast amount of data. Big data are mainly the product of the datafication process [39], which implies the transformation of economic, social and, in general, individual life into data. Prosumerism [36], one of the main features of web 2.0, refers to the fact that most of these big data are produced by web users: the so-called "user-generated content" (UGC). UGC is the content (i.e., text, images, video) posted and freely shared by users on blogs, forums, social media, wikis, and so on.

This kind of big data can sometimes convey geographical information (GI), which has evolved with Web 2.0. Other kinds of data contain GI, such as mobile phone data, and for this reason, Niu and Silva [46] to define them all proposed the unique concept of "crowdsourced geographic information" (CGI).

There are at least three ways to obtain geographic information from the data [42]. The first is the geotagging process which occurs when spatial coordinates are assigned to the data. The other two ways to extract spatial information are geocoding and geoparsing. Geocoding refers to the transformation of a well-formed textual representation of an address into a valid spatial representation, while geoparsing is the extraction of geographic information from unstructured free texts.

Once this data has been collected, it is possible to classify them according to two criteria: a) voluntariness in content creation and b) the role played by the territory in the content. By combining these two criteria, it is possible to distinguish at least four types of data (Cfr. Table 1).

Table 1. *Type of CGI.*

		Voluntariness	
		Volunteer	Non-voluntary
Territory's role	Primary	territory's information produced and shared by users	information collected by navigation apps while using them
	Attachment	thoughts and opinions shared by users containing geographical references	information about the user's location during daily activities

In the first type, the territory is the focus of the voluntary user activity. Goodchild [20] has called this data "volunteered geographic information" highlighting that individuals intentionally add information regarding the territory. Platforms like Wikimapia, Google Earth and photo-sharing web applications like Flickr, Panoramio and Picasa Web are large databases of this kind of data provided by users that share information about a specific location or specific areas. This kind of geographical information has been called "volunteered geographic information" to stress that the user was actively involved in the information creation process.

In the second kind of data, the content is created voluntarily by the user but has no direct reference to the territory. In such cases, spreading geographic information is not the final purpose of content production [52]. For example, a Twitter or Facebook user who automatically adds his or her position when sharing a thought or opinion. It is important to emphasise that is possible to find also first-type data, such as Facebook' check-in' data on these platforms.

In the third and fourth kinds of data, there is no intention to share any content with others. The third kind of data comes from a platform or app where space plays a central role in the user experience; for instance, google maps that help the user find their way to a particular place. The user, through these apps, gives information about his or her location and movements. This data is collected in the background and serves many purposes, including commercial ones.

In the fourth type of geographic data, the spatial dimension plays no role, and the user is unaware that data on his or her

location is being collected.

Most telephone networks generate Call Detail Records (CDRs), i.e., data records produced by a telephone exchange that document the details of a call or SMS passed through the device. CDR data are suited to tracking the entire population with relatively high spatial accuracy [4, 58].

Moreover, some apps, that do not offer geographic services, also collect data on the user's location, requesting the permission to detect it with the promise of improving the service.

The spread of geographical data of users gave birth to the spatial turn in media studies and the media turn in geographical studies [55]. In particular, the availability of the first and second types of CGI has made it possible to study territory's influence on social phenomena that have communicative implications.

3. Methodology

To address the aim of this research, a systematic literature review (SLR) was carried out during a temporal range that goes from September 2022 to November 2022. The SLR was conducted using PRISMA methodology [44, 45], a rigorous procedure to search, select, and analyse the findings from the literature based on the objective of the study.

According to other systematic reviews and to the PRISMA, the steps of identification, screening, eligibility, and inclusion [44] were followed. This plan guaranteed the methodological accuracy, replicability, and transparency of research [57].

3.1. Identification

The research was carried out in the academic research databases SCOPUS and Web of Science Core Collection (WOS) databases, choose for their extensiveness and relevance in social sciences [47]. Research in the different databases focuses on the following research string: ((("ecological analysis" OR "territorial" OR "physical space" OR "urban space" OR "spatial characteristics" OR "spatial properties" OR "spatial effect" OR "neighbourhood effect" OR "neighborhood effect" OR "contextual effects" OR "environmental effects" OR "spatial analysis") AND ("user generated content" OR "crowdsourced geographic information" OR "Volunteered Geographic Information" OR "social media" OR "social networks" OR "weibo" OR "facebook" OR "twitter" OR "tiktok" OR "instagram" OR "reddit" OR "wechat" OR "whatsup" OR "foursquare" OR "flickr" OR "kuaishou" OR "pinterest" OR "douyin" OR "panoramio" OR "wikimapia" OR "picasa"))). The keywords, connected with the "AND" and "OR" Boolean operators, needed to appear in the title, in the keywords or in the abstracts of the papers to ensure a comprehensive search. As can be seen, the search string is divided into two parts. The first contains references to concepts relating to spatial analysis (urban space; territorial; etc.) and NE (neighbourhood effect; contextual effects; environmental effects,...). The second part of the string concerns social media through generic references (e.g., social media; social networks) and punctual ones (e.g., Facebook, Twitter, etc.) references. 2560 records were produced from Scopus and 1378 from WOS (*cfr.* Table 2).

Table 2. *The process of data collection.*

Database	Keywords	Records
Scopus	TITLE-ABS-KEY: ((("ecological analysis" OR "territorial" OR "physical space" OR "urban space" OR "spatial characteristics" OR "spatial properties" OR "spatial effect" OR "neighbourhood effect" OR "neighborhood effect" " OR "contextual effects" OR "environmental effects" OR "spatial analysis") AND ("user generated content" OR "crowdsourced geographic information" OR "Volunteered Geographic Information" OR "social media" OR "social networks" OR "weibo" OR "facebook" OR "twitter" OR "tiktok" OR "instagram" OR "reddit" OR "wechat" OR "whatsup" OR "foursquare" OR "flickr" OR "kuaishou" OR "pinterest" OR "douyin" OR "panoramio" OR "wikimapia" OR "picasa")))	2560
WOS	TITLE-ABS-KEY: ((("ecological analysis" OR "territorial" OR "physical space" OR "urban space" OR "spatial characteristics" OR "spatial properties" OR "spatial effect" OR "neighbourhood effect" OR "neighborhood effect" "contextual effects" OR "environmental effects" OR "spatial analysis") AND ("user generated content" OR "crowdsourced geographic information" OR "Volunteered Geographic Information" OR "social media" OR "social networks" OR "weibo" OR "facebook" OR "twitter" OR "tiktok" OR "instagram" OR "reddit" OR "wechat" OR "whatsup" OR "foursquare" OR "flickr" OR "kuaishou" OR "pinterest" OR "douyin" OR "panoramio" OR "wikimapia" OR "picasa")))	1378

3.2. Screening and Eligibility

After the first step, the inclusion and exclusion criteria (*cfr.* Table 3) to obtain the relevant literature in database [7] were set. Regarding the type of document reviews, editorial, conference papers, "in press" papers, book chapters and working papers were excluded, while articles published in international and national peer-reviewed journals in English language belonging to the research field of social science were included in the analysis.

Table 3. *Inclusion Criteria.*

Criteria	Description
Document Type	Peer-reviewed journal articles
Language	English
Research Domain	Social Science
Timeline	Until November 2022

It is important to highlight that no temporal range was set, neither a minimum number of citations were not included because of the novelty of the topic under study. On this basis, 975 records from Scopus and 521 from Web of Science were obtained. 294 out of 1496 records, obtained from Scopus and Wos, were rejected as duplicates. Based on the reading of 1202 abstracts, 949 papers were deleted, because of their low pertinence to research aims and review objectives. Most of the delete records contained the keyword "network analysis" referred to the methodology used within the work and was not pertinent with the research aim.

3.3. Inclusion

In the last phase of the PRISMA methodology, after reading the full texts of all 253 remaining articles, 194 publications were included in the review process to individuate the main themes related to the CGI through the bibliometric analysis. Then, to address the question related to the NE approaches, a content analysis on 47 articles was performed (Figure 1).
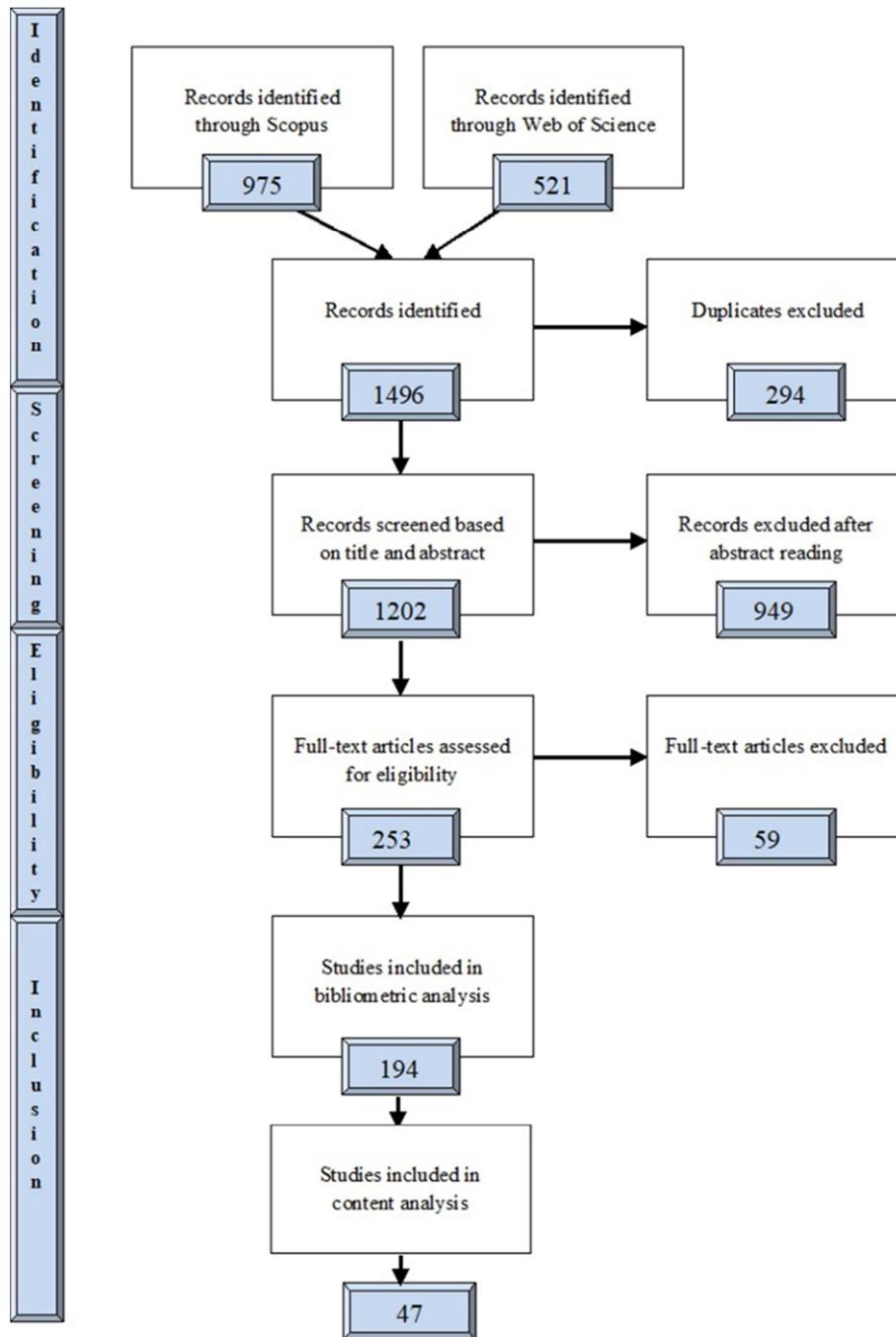
Figure 1. The assessment and selection of contributions: PRISMA flow diagram.

4. Findings

4.1. The CGI Scientific Production

Vosviewer, a program developed primarily for bibliometric analysis, was used to identify the network of central cores of the scientific production related to CGI. The article's keywords were chosen as reliable proxies for the fields and the topics studied in the articles. The suitable keywords for the network were selected according to a minimum occurrence threshold (4). Of the 1313 keywords, less than 8% met the threshold. In addition, keywords related to the spatial statistical techniques used within the papers were removed.

The occurrence threshold of 4, by excluding the less popular keywords, brought to light the modal cores of the scientific production using CGIs (RQ1). As shown in Figure 1, there are four thematic cores identified. In the first two, CGI are mostly related with population's movements in space, time and among various urban activities.

The two clusters refer to (a) tourism and (b) urban planning.

Regarding the first cluster, its component keywords are connected to the various aspects of the tourism phenomenon:

the tourist flows management; the identification of spatial patterns that explain their distribution in the city [60]; the discovery of emerging places from the touristic point of view [34] the link between the places' heritage and the number of tourists [29], and so on.

There is a thriving literature using CGI focused on tourism, which is very valuable due to the lack of up-to-date databases on the phenomenon. Moreover, in this cluster appears the social Flickr [29, 60], a platform where users can upload their photos, especially regarding visited places. This kind of photo are often used as trace of the people's passage, making possible the study of the phenomenon.

In the second cluster, that collect the scientific production on urban planning, there are used similar data to the previous cluster analysed. From the analysis of the works emerges that

the topic of urban planning, especially within the smart city, is gaining increasing interest. This cluster is characterised by contributions based on human activities and mobility, and its ecological unit of analysis is the city. For example, some authors try to identify the factors that are be able to predict which places could develop more urban vibrancy [58] while others assess the accessibility of urban spaces and what are the points of interest [24]. Moreover, starting from points of interest, according to Yin and colleagues [65] could be delineate/define urban boundaries through the analysis of human activity rather than through administrative limits. The third and fourth clusters are included research that considers, as well as the spatial aspect, the communicative aspect of CGI.
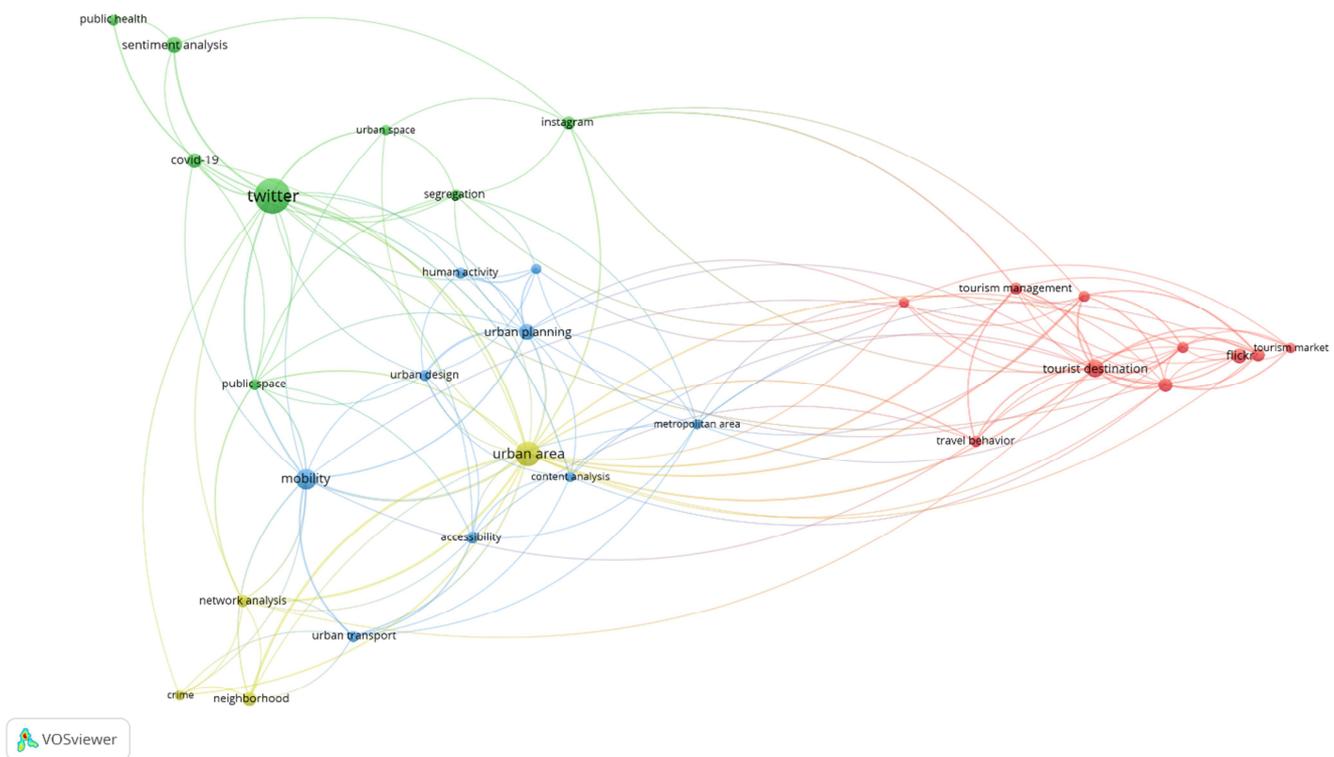
Figure 2. Main Topi.

In fact, in the third cluster, there is the keyword sentiment analysis, which is a technique used to analyse the content shared by the user. Accordingly, this cluster's core is not a particular theme but the data sources, such as Instagram and Twitter. In particular, the latter social turned out to be a precious source for social researchers because it represents a database where it is possible to find individuals' opinions, thoughts, and attitudes on many social phenomena. In addition, given the fact that Twitter's public APIs provide easy access to the platform's data, social media monitoring systems usually start from there. Moreover, in this cluster, there is a strong influence of the COVID-19 pandemic. The scientific production related to the COVID-19 pandemic is impressive. The collected studies have used different analytic techniques, including spatial-temporal analysis, to investigate

the relationship between the spread of the pandemic in territories and the citizens' sentiment toward COVID-19 [2] or vaccine [64]. The pandemic, with the resulting lockdown, affected every aspect of social life, including the spatial arrangement; so, it is possible to find studies that combine the theme of space use - pre- and post-pandemic - with segregation [33]. Furthermore, segregation, as topic emerged in this cluster, is analysed both related to the COVID-19 and through its constitutive aspects, such as the language segregation [59]. The core of the last cluster is the concept of neighbourhood, understood as an ecological unit. Regarding the neighbourhood the most frequent studies deal with crime [9] and social networks [62].

In particular, the works on social networks regard the relationship between online networks and the type of

neighbourhood [18] and the online network and the network in physical space [62].

Although it plays a role in this cluster, the neighbourhood concept turns out to be outside the centre of the network as it was possible to imagine. There are two reasons behind this result. The first is terminological; as seen in the first paragraph, there are different concepts to identify the influence of the territory on the investigated phenomenon. The other reason is related to the conception of space that prevails in the works and this is the other result that emerges from the overall reading of the literature.

So, it is possible to divide the analysed articles into two macro-areas according to the type of relation with the territory that emerges from the analysed phenomenon. The first area is characterised by a geographical approach inasmuch as it deals with the analyses of the spatial distribution of the analysed phenomenon. The second area, on the other hand, deal with the influences that territory's characteristics can have on the phenomenon analysed. Moreover, in this area, it is possible to find works that focus on neighbourhood characteristics, that are the 25% of the articles that built the network. Thus, these contributions were analysed to address the last three RQs presented above.

4.2. The Sources Used

Three findings emerge from the analyses of the sources used in the articles examined (RQ2). The first one regards the geography of social media used as a database. As seen in table 4, among the different social media, it is possible to find Weibo, the most used social media for studying the Chinese context, in which, due to governmental censorship, many of the social media used in the West are blocked.

A social platform can be a good data source only if it has an adequate number of users in its area of reference. For example, Twitter could be a handy data source in countries such as Spain, England, or the United States, where at least 40 per cent of the network users use it monthly, while in Germany, the users are only the 20 per cent. However, the spread of a social platform among the population does not guarantee its representativeness. The social's platform population characteristics are not the same as the overall population in terms of sociodemographic characteristics and spatial distribution [3, 21, 27]. Urban areas, for example, are often over-represented compared to inland areas [11]. Since each social platform has its target audience, the bias between them is not equal. Using geolocated data and techniques for retrieving user sociodemographic information could be useful in dealing with these biases. The second result that emerges, looking at the table 4, is that Twitter is the social media most used as a data source. This is undoubtedly related to the fact that it has a very high number of global active users (436 million - [22]), and in some countries, it is used by almost half of the network's users. Since it is a micro-blogging platform, it is possible to retrieve textual material from it. Similarly, through the network analysis it is possible to analyse the network structure between concepts and subjects. The main reason for this success lies in the fact that Twitter,

through the API, allows free downloading of user activity data; this characteristic is crucial in the current search environment defined as post-API [15] because the proprietary closure [38]. Anyway, like Twitter also Instagram, although with some limitations, allows API access. The frequency distribution results (table 4) show that Facebook - the most used social in the world (almost 3 billion active users - [22]) - is rarely used in research. That is clearly due to data restrictions that have become even tighter since the Cambridge Analytica case. The last result concerns the use of multiple CGI sources. For example, 10% of the papers consider two CGI sources. It is important to specify that, beyond CGI sources, these research are multi-source since CGI data are usually related to other types of data drawn from pre-existing databases such as the census database. The analysis of the sources shows that data policy, as well as government and citizen policy, considerably influences both 1) the actions of researchers in this field when it places circumventable barriers to the use of data; and 2) most parts of the research regarding social media. Currently, Twitter is social with minor restrictions and, therefore, most used for spatial media analysis, even though only a small percentage of tweets [66] contain the geographic tag. It is considered equally interesting that Facebook's contribution to this strand of analysis is almost irrelevant.

Table 4. Main sources used.

Source	Freq
Facebook	2
Foursquare	3
Instagram	3
Weibo	5
Twitter	40
Tot. sources	53
Tot. Articles	47

4.3. How to Relate CGI to Territory

Working on the Twitter platform means accessing the textual content produced by users. It is no by chance that about a third of the articles analysed, take into account the textual content of tweets Despite that, in an equally large proportion of articles is analysed the simple trace in space left by users. For example, in studies concerned the Jacobsian concept of Urban Vibrancy [25], the position in the space is detected using the user's number of access into shops or in public places as indicators [32, 58]. Based on this information, Adelfio and colleagues [1] identify an inverse relationship between the economic well-being of the neighbourhood and the vitality of public spaces. In a smaller proportion of articles, instead, the focus is on geolocalized online networks, that are useful proxy to study the social vitality of neighbourhoods. For example, Gibbons and colleagues [19] study the density of location-based interactions between gentrifying and non-gentrifying neighbourhoods. As shown by the analysis, the informative contribution of CGI to social phenomena studies is various. Working on the textual content, more than the physical trace, ensures a better flexibility to the researcher. From the users'

verbal expressions, in fact, can be derived information about a lot of social phenomena. So, whatever information is gleaned from the CGIs, it is always interesting understand how this content is combined with both the characteristics of the in-vivo world and the territory.

Table 5. *Territorial variable used.*

Variables	Freq.
Real estate market	5
Local economy	6
Infrastructure	10
Medial	11
Physical/ Positional	11
Socio-demographic	23

Table 5 shows the groups of independent variables used in the articles (RQ3). There have been identified seven modes: socio-demographic; media; infrastructural; physical/location; local economy; real estate market, and one defined as 'other' which includes less used aspects, such as the characteristics of the natural environment. Regarding the category 'media', the analysis shows that the 80% of the studies constructs its dependent variable from the information from CGIs; while in a smaller proportion of studies (less than 20%), the dependent variable is constructed on information from classical databases. In these articles, information from CGIs, the so-called 'medial variables', intervene in the analysis as independent variables. The medial variable often consists of textual content, as it is possible to see in the research by Ristea and colleagues [49] who include the number of tweets with violent content in their regression model to explain the number of violent acts. This research seems to embrace the conceptualization of space borrowed from Lazarsfeld's studies, where the focus is on the communicative aspect of the neighbourhood. Regarding Galster's conceptualisation [16], considering that textual content may convey values and beliefs, it is considered that the "media variable" forms the background to Social-interactive Mechanisms. This type of mechanism characterizes also studies that consider socio-demographic-economic explanatory variables, such as the percentage of university graduates, the percentage of foreign population, the percentage of employed people and so on. These are the most frequently used variables, and half of the articles analysed use them. On the other hand, even in ecological studies, population characteristics are often used to explain multiple social phenomena. In terms of usage, after the socio-demographic variables, it is possible to find the already mentioned medial variables, then the physical location and infrastructure variables (i.e., metro and/or roads and their physical characteristics, the quality of spaces, parks, etc.). These variables refer to environmental mechanisms. In this group of mechanisms there are included also the variables concerning the real estate market, used in five articles. With the physical/location variables, authors identify the metric characteristics of space both in an absolute sense, such as the m^2 of the ecological unit investigated [23] and in a relative sense, such as the distance of a given area from the city centre, an infrastructure, or a service [17]. These

characteristics make it possible to identify the geographical mechanisms that refer to the various spatial aspects of the investigated ecological unit. Finally, there is the variable related to the local economy, which is used in six articles and basically relates to business activity characteristics. This variable recalls the institutional mechanisms. As emerged above, the category "other" includes variables with an unremarkable frequency, such as the characteristics of the natural environment. It is unlikely to uniquely connect this category to a mechanism because the variables included in this mode recall both institutional and environmental mechanisms. The sum of the frequencies used gives a higher value than the articles examined. That is because some articles in their explanatory models take more than one variable, and consequently more mechanisms, into account at the same time. All the articles analysed, as seen, relate the phenomenon to territory's characteristics, i.e., socio-demographic, institutional, media etc. It is clear that this is not a proper result but the product of the paper selection, which excluded the ones apart from the tradition of NE studies. Findings also shows that all the articles use a quantitative approach (RQ4) and none use qualitative or mixed methods. Due to their amount, big data hardly can investigate into purely qualitative studies; however, in-depth studies of the emerging results to corroborate the mechanisms identified in the analysis could enrich the debate. In the articles, it is possible to appreciates varied techniques that implement the spatial dimension in statistical analysis to identify spatial patterns.

5. Discussion and Research Agenda

This study carried on a SLR to show that there is a new, although uncertain, frontier to explore for NE constituted by CGIs. *New Frontier's assessment:* as seen in the analysis of thematic cores, CGIs represent a precious information resource for territory management. It is not by chance that one of the pillars of the smart city concept is monitoring the territory through geolocalized data. CGIs can also be very useful in the study of NE. In addition to physical traces, it is possible to infer information on multiple phenomena through textual and network analysis. However, this possibility is closely linked to the actual availability of the information on the territory. In fact, many of the studies reviewed concern contexts where devices are used to detect CGIs. Moreover, the spread of social platforms is different among population groups or territories, making the possibility of generalising results very distant. This would be difficult also since a large part of the CGIs does not contain socio-demographic information, that maybe can be obtained in the future thanks to the development of predisposed algorithms. These limitations should not be neglected because they represent the work that can be done on the topic, to cross the new frontier Thanks to the review of the work, it was possible to see, for instance, that the content of tweets varies according to the territory's characteristics, or the online networks vary according to the characteristics of the neighbourhood and/or

spatial proximity. Therefore, it is clear that CGI can be useful in the study of territory and of NE, and to deepened over time the relationship between the online world and the in-vivo world. *New frontier's exploration*: the production of NE that includes the use of CGIs is still in its infancy, and this is confirmed by the number of articles identified for review, which is less than 50. In the articles analysed, a quantitative approach to the study of NE prevails, and this is clearly in line with the fact that CGI is Big Data. Qualitative studies concerning the online and in-vivo world do exist, but they mainly concern how the online contributes to changing the interpretation of the in-vivo world. Integrating a qualitative step into NE studies could cost a lot, especially if this phase could only carry on through traditional tools. However, besides quantitative analyses, it is also possible to carry out qualitative analyses on data from social platforms in a mixed-method perspective. This type of analysis requires fundamental theoretical skills and digital and data analysis skills. Skills that are easier to find in multidisciplinary groups. Thus, the NE approach could be digitize and broaden by identifying, for example, new reference concepts. The analysis of the variables and the mechanisms underlying it revealed a full awareness of the possibility that territory and neighbourhood have to influence social phenomena. In addition to the classic socio-anagraphic variables, less obvious variables, such as infrastructural and/or institutional ones, were considered. This indicates research specifications that reflect different conceptions of space. This is the case of the inclusion of medial variables, that seems to be in line with the concept of space borrowed from Lazarsfeld's American studies. This conceptualization completes the more classical one derived from Durkheim's seminal concepts. Importing a theoretical space conception, whether American or French derivation, within studies concerning CGI is fundamental to avoid naïve analytical approaches to territorial phenomena. On the other hand, the inclusion of CGIs could advance studies on NE. This is the case of the concept of triple neighbourhood disadvantage, which imports the digital dimension into the classical concept of disadvantage through the use of in- and outgoing networks. *New frontier's uncertainties:* Currently, a large part of the CGI used comes from private companies that make their data accessible. This not only could influence the research directions but also makes them uncertain. Working on pre-established and locked databases, could force researchers to identify the own research questions upstream, so they could be addressed with the available data. On the other hand, it is important that these data can be used as valid proxies for social phenomena, as shown in the above review. A particularly critical aspect regarding the new frontier is the possibility for private individuals to shut down valuable channels for knowledge of the social world, thwarting, in this way, also the previous cognitive efforts. Algorithms for analysing the textual content of tweets - also with a geolocalization point of view (the geoparsing algorithms) - have been developed. However, if the new owner of Twitter, Elon Musk, 'shuttered down' Twitter's API legitimately, these

cognitive efforts would be lost as well as the possibility of producing knowledge through that channel. Compared to the past, now there are private organizations that possess an essential amount of information and data on the social world and its dynamics. So, it become more important that traditional public data sources should catch up with the times by integrating new data sources (i.e., CGIs) and making them available to researchers and to various stakeholders. This would ensure continuity and increased knowledge capacity.

Thus, the analysis of the social research literature carried on through the thematic analysis, the analysis of the sources, the one regarding the variables and the approaches' analysis focuses on CGIs led to the identification of some gaps and future research directions.

Research direction 1: Implementing mixed methods approach

The review revealed that there is the need for implementing in-depth qualitative techniques to provide further robustness to the results and possibly corroborate the agent mechanisms. For this reason, it might be helpful to implement mixed methods approaches in order to analyse digital data [12].

Research direction 2: Increase of multi- and inter-disciplinarity

The social research on CGIs is dynamic and is developing on fragmented disciplinary lines. In particular, the review, through the analysis of existing works, has shown the need of interconnections among different fields of knowledge [14], for better manage the topic by looking at the totality of the phenomenon.

Research direction 3: public production of CGI

In order to explore the above research directions, however, on the one hand, the problems of demographic and territorial representativeness, and on the other the actual availability of data, will have to be addressed. From the review emerged an important role for public organizations and governments in providing accessible online information for scientist and a greater public bodies' and researchers' involvement in producing CGIs could be a key to future studies.

6. Conclusion

This paper provides a synthesis of the current conceptual and empirical literature on CGIs. The literature review revealed that CGIs are progressively being implemented in N.E. studies. The first theoretical contribution of this study pertains to the demonstration of possible relations between the online world and the in-vivo world, these works, in fact, have shown the existence of a new frontier for the study of N.E. The second implication of the review lies in the harmonisation of the existing knowledge in four macro-areas that underlines which are the main stressed topics in literature. Moreover, from a theoretical point of view the concept of territory borrowed from the approach and present in the articles analysed, even if not in an explicit form, may be helpful to those who are preparing to analyse CGI. Eventually, one of the most important implications of the review, in fact, is the

development of a research agenda, highlighting that the approach is promising because it opens new investigation scenarios and lends itself to developing new analytical and theoretical concepts. Regarding the practical implications, this study can contribute to a general understanding of the phenomenon for policymakers who wants to address better the territory management, highlighting the possibility to create, through CGI, data strategies to transform the government and public services, making them more citizen-centric, responsive, accountable, and transparent. Moreover, the work shows the need for policymaker to create public dataset of essential data for research useful also for sustainable urban planning and development.

References

[1] Adelfio, M., Serrano-Estrada, L., Martí-Ciriquián, P., Kain, J. H. & Stenberg, J. (2020). Social activity in Gothenburg's intermediate city: Mapping third places through social media data. *Applied Spatial Analysis and Policy*. 13, 985-1017. doi: 10.1007/s12061-020-09338-3.

[2] Alagha, I. (2021). Topic modeling and sentiment analysis of Twitter discussions on COVID-19 from spatial and temporal perspectives. *Journal of Information Science Theory and Practice*. 9, 35-53. doi: 10.1633/JISTaP.2021.9.1.3.

[3] Ballatore, A. & De Sabbata, S. (2020). Los Angeles as a digital place: The geographies of user-generated content. *Transactions in GIS*. 24, 880-902. doi: 10.1111/tgis.12600.

[4] Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J. & Ratti, C. (2011, October). The connected states of america: Quantifying social radii of influence. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (pp. 223-230). IEEE.

[5] Catalano, R. & Pickett, K. E. (2003). "A Taxonomy of Research Concerned with place and health", in The handbook of social studies in health and medicine, ed. Albrecht G. L., Fitzpatrick R., Scrimshaw S. C. (Thousand Oaks, CA: Sage Publication), 64-83.

[6] Cavan, R. S. (1983). The Chicago school of sociology, 1918-1933. *Urban life*. 11, 407-420. doi: 10.1177/0098303983011004003.

[7] Cooper H. (1998) Synthesizing Research: A Guide for Literature Reviews, 3rd edn. Sage Publications, Thousand Oaks, CA.

[8] Cox, K. R. (1969). The voting decision in a spatial context. *Progress in geography*. 1, 81-117.

[9] Cutchin, M. P., Eschbach, K., Mair, C. A., Ju, H. & Goodwin, J. S. (2011). The socio-spatial neighborhood estimation method: an approach to operationalizing the neighborhood concept. *Health & place*. 17, 1113-1121. doi: 10.1016/j.healthplace.2011.05.011.

[10] Davis, J. A. (1961). Compositional effects, role systems, and the survival of small discussion groups. *Public Opinion Quarterly*. 25, 574-584. doi: 10.1086/267053.

[11] De Falco, A., De Falco, C. C. & Ferracci, M. (2022). Geo-Social Media and Socio-Territorial Distribution: A Study on the Italian Case. *Italian Sociological Review*. 12, 685-708. doi: 10.13136/isr.v12i7S.577.

[12] De Falco, C. C., Punziano, G. & Trezza, D. (2021). A mixed content analysis design in the study of the Italian perception of COVID-19 on Twitter. *Athens Journal of Social Sciences*. 8, 191-210. doi: 10.30958/ajss.8-3-3.

[13] de Vos, S. (1998). The analysis of compositional effects as exemplified by the study of elections. *GeoJournal*. 44, 43–49. doi: 10.1023/A:1006820202876.

[14] Emani, C. K., Cullot, N. & Nicolle, C. (2015). Understandable big data: a survey. *Computer science review*. 17, 70-81.

[15] Freelon, D. (2018). Computational research in the post-API age. *Political Communication*. 35, 665-668. doi: 10.1080/10584609.2018.1477506.

[16] Galster, G. C. (2012). "The mechanism (s) of neighbourhood effects: Theory, evidence, and policy implications", in Neighbourhood effects research: New perspectives, (Dordrecht: Springer), 23-56.

[17] Ghosh, D. & Guha, R. (2013). What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and geographic information science*. 40, 90-102. doi: 10.1080/15230406.2013.776210.

[18] Gibbons, J. (2022). Why can't we be friends? The underlying hierarchy of social connections between diverse neighborhoods. *Applied Spatial Analysis and Policy*. 15, 1339–1365. doi: 10.1007/s12061-022-09453-3.

[19] Gibbons, J., Nara, A. & Appleyard, B. (2018). Exploring the imprint of social media networks on neighborhood community through the lens of gentrification. *Environment and Planning B: Urban Analytics and City Science*. 45, 470-488. doi: 10.1177/2399808317728289.

[20] Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*. 69, 211-221. doi: 10.1007/s10708-007-9111-y.

[21] Haffner, M. (2018). A spatial analysis of non-English Twitter activity in Houston, TX. *Transactions in GIS*. 22, 913-929. doi: 10.1111/tgis.12335.

[22] Hootsuite (2022). We are social - Digital 2022 Global Overview Report.

[23] Hu, L., Li, Z. & Ye, X. (2020). Delineating and modeling activity space using geotagged social media data. *Cartography and Geographic Information Science*. 47, 277-288. doi: 10.1080/15230406.2019.1705187.

[24] Iranmanesh, A. & Atun, R. A. (2018). Exploring the spatial distribution of geo-tagged Twitter feeds via street-centrality measures. *Urban Design International*. 23, 293-306. doi: 10.1057/s41289-018-0073-0.

[25] Jacobs, J. (1961). *The death and life of great American cities* (reprint). New York: Random House.

[26] Jargowsky, P. A. (2005). The ecological fallacy. *Encyclopedia of social measurement*. 1, 574-584.

[27] Jiang, Y., Li, Z. & Ye, X. (2019). Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and geographic information science*. 46, 228-242. doi: 10.1080/15230406.2018.1434834.

[28] Johnston, R. (1979). Class conflict and electoral geography. *Antipode*. *11* (3), 36-43.

[29] Karayazi, S. S., Dane, G. & Vries, B. D. (2021). Utilizing urban geospatial data to Understand heritage attractiveness in Amsterdam. *ISPRS International Journal of Geo-Information*. 10, 198. doi: 10.3390/ijgi10040198.

[30] King, G., Tanner, M. A. & Rosen, O. (2004). *Ecological inference: New methodological strategies*. Cambridge: Cambridge University Press.

[31] Knies, G., Melo, P. C. & Zhang, M. (2021). Neighbourhood deprivation, life satisfaction and earnings: Comparative analyses of neighbourhood effects at bespoke scales. *Urban Studies*. 58, 2640-2659. doi: 10.1177/0042098020956930.

[32] Lang, W., Lang, H., Hui, E. C., Chen, T., Wu, J. & Jahre, M. (2022). Measuring urban vibrancy of neighborhood performance using social media data in Oslo, Norway. *Cities*. 131, 103908. doi: 10.1016/j.cities.2022.103908.

[33] Legeby, A., Koch, D., Duarte, F., Heine, C., Benson, T., Fugiglando, U. & Ratti, C. (2022). New urban habits in Stockholm following COVID-19. *Urban Studies*. https://doi.org/10.1177/00420980211070677

[34] Li, Y., Xie, J., Gao, X. & Law, A. (2021). A Method of selecting potential development regions based on GPS and social network models–from the perspective of tourist behavior. *Asia Pacific Journal of Tourism Research*. 26, 183-199. doi: 10.1080/10941665.2018.1515092.

[35] Liebst, L. S. Ä. (2016). Reassembling Durkheimian sociology of space. In Spatial Cultures (pp. 246-256). Routledge.

[36] Lupton, D. (2014). *Digital sociology*. London: Routledge.

[37] Macintyre, S., Maciver, S. & Sooman, A. (1993). Area, class and health: should we be focusing on places or people? *Journal of social policy*. 22, 213-234. doi: 10.1017/S0047279400019310.

[38] Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), Debates in the Digital Humanities. Minneapolis, MN: The University of Minnesota Press. http://manovich.net/index.php/projects/trending-the-promises-and-the-challenges-of-big-social-data

[39] Mejias, U. A. & Couldry, N. (2019). Datafication. *Internet Policy Review*. 8, 261-278. doi: 10.14763/2019.4.1428.

[40] Mela, A. (2006). *Sociologia delle città*. Roma: Carocci.

[41] Menzel, H. (1950). Comment on Robinson's "ecological correlations and the behavior of individuals". *American Sociological Review*. 15, 674-674.

[42] Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S. & Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS)*. 36, 1-27. doi: 10.1145/3202662.

[43] Miller, W. L. (1977). *Electoral dynamics in Britain since 1918*. London: Palgrave Macmillan.

[44] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*. 151 (4), 264-269.

[45] Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews*. 4 (1), 1-9.

[46] Niu, H. & Silva, E. A. (2020). Crowdsourced data mining for urban activity: Review of data sources, applications, and methods. *Journal of Urban Planning and Development*. 146, 04020007. doi: 10.1061/(ASCE)UP.1943-5444.0000566.

[47] Norris, M. & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of informetrics*. 1 (2), 161-169.

[48] Oakes, J. M., Andrade, K. E., Biyoow, I. M. & Cowan, L. T. (2015). Twenty years of neighborhood effect research: an assessment. *Current epidemiology reports*. 2, 80-87. doi: 10.1007/s40471-015-0035-7.

[49] Ristea, A., Al Boni, M., Resch, B., Gerber, M. S. & Leitner, M. (2020). Spatial crime distribution and prediction for sporting events using social media. *International Journal of Geographical Information Science*. 34, 1708-1739. doi: 10.1080/13658816.2020.1719495.

[50] Sampson, R. J., Morenoff, J. D. & Gannon-Rowley, T. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual review of sociology*. 177-187. Doi: https://doi.org/10.1146/annurev.soc.28.110601.141114.

[51] Settersten, R. A. (2001). "Community Organization and the Life Course", in International Encyclopedia of the Social & Behavioral Sciences, ed. N. J. Smelser, P. B. Baltes (Oxford, UK: Pergamon Press), 2367-2371.

[52] Stefanidis, A., Crooks, A. & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*. 78: 2, 319–338. Doi: https://doi.org/10.1007/s10708-011-9438-2.

[53] Stein, R. E. (2014). Neighborhood scale and collective efficacy: Does size matter? *Sociology Compass*. 8: 2, 119-128. https://doi.org/10.1111/soc4.12127

[54] Strassoldo, R. (1990). TheSociology of Space, Culture-Space-History: Proceedings 11th International Conference of the IAPS, Ankara, Turkey.

[55] Thielmann, T. (2010). Locative media and mediated localities. *Aether: the journal of media geography*. 5, 1-17.

[56] Tingsten, H. (1937). *Political Behavior; Studies in Election Statistics*. London: PS King and Son.

[57] Tranfield, D., Denyer, D. & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*. 14 (3), 207-222.

[58] Tu, W., Zhu, T., Xia, J., Zhou, Y., Lai, Y., Jiang, J. & Li, Q. (2020). Portraying the spatial dynamics of urban vibrancy using multisource urban big data. *Computers, Environment and Urban Systems*. *80*, 101428. Doi: 10.1016/j.compenvurbsys.2019.101428.

[59] Väisänen, T., Järv, O., Toivonen, T. & Hiippala, T. (2022). Mapping urban linguistic diversity with social media and population register data. *Computers, Environment and Urban Systems*. 97, 101857. Doi: https://doi.org/10.1016/j.compenvurbsys.2022.101857.

[60] Valls, F. & Roca, J. (2021). Visualizing digital traces for sustainable urban management: mapping tourism activity on the virtual public space. *Sustainability*. *13*: 6, 3159. Doi: https://doi.org/10.3390/su13063159.

[61] van Ham, M. & Manley, D. (2012). Neighbourhood Effects Research at a Crossroads. Ten Challenges for Future Research Introduction. *Environment and Planning A*. 44: 12, 2787-2793 Doi: https://doi.org/10.1068/a45439.

[62] Wang, Z., Ye, X., Lee, J., Chang, X., Liu, H. & Li, Q. (2018). A spatial econometric modeling of online social interactions using microblogs. *Computers, Environment and Urban Systems*. *70*, 53-58. Doi: https://doi.org/10.1016/j.compenvurbsys.2018.02.001.

[63] Wilson W.J. (1987). The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. Chicago: University of Chicago Press.

[64] Yin, F., Crooks, A. & Yin, L. (2022). Information propagation on cyber, relational and physical spaces about COVID-19 vaccine: Using social media and splatial framework. *Computers, Environment and Urban Systems*. *98*, 101887. Doi: 10.1016/j.compenvurbsys.2022.101887.

[65] Yin, J., Soliman, A., Yin, D. & Wang, S. (2017). Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science*. *31*: 7, 1293-1313. Doi: https://doi.org/10.1080/13658816.2017.1282615.

[66] Zola, P., Cortez, P. & Carpita, M. (2019). Twitter user geolocation using web country noun searches. *Decision Support Systems*. *120*, 50-59. Doi: https://doi.org/10.1016/j.dss.2019.03.006.