Defining a Micro-Geodemographic *Natural Area* with Street-Network Topology

Peter Prescott

2021

Abstract

Geodemographic typologies provide a powerful way of helping us understand the complexity of human social reality, but they have been criticized for being theoretically ungrounded, ontologically nebulous, and ethically problematic. In this paper I show how a simple theory of neighbourly relational availability gives rise to a socially-grounded geographically definition of neighbourhood based not on Euclidean geometric distance, but on topological street-network connection. I demonstrate how the recently-released Ordnance Survey UPRN dataset can be combined with other Open Data products to operationalize this definition. And I apply it, with free and open-source code, to the whole of mainland Britain, in less than thirty-eight minutes.

Contents

List of Tables

List of Figures

1. Introduction

We are still only in the early days of a digital revolution that has already begun to transform every aspect of human life, and will continue to do so in ways that can hardly yet be imagined, let alone quantifiably predicted (Batty, 2021). Paradoxically, one of the most powerful drivers of this general trend of unpredictable change is the increasing power of sophisticated algorithmic technologies to extract specific predictive insights concerning some sufficiently well-defined question (Openshaw & Openshaw, 1977; Russell & Norvig, 2016). In a powerful feedback loop, the more complex our contemporary social context becomes, the more necessary become those tools which enable informed decision-making amid an unprecedented deluge of (potentially) relevant data (Kitchin, 2014). One such tool is *geodemographics* (Harris et al., 2005; Webber & Burrows, 2018).

Geodemographic analysis applies unsupervised machine learning to the demographic data associated with geographic areas, thus enabling the reduction of the complex multidimensional reality of human society to a more manageable number of statistical types. Having been identified algorithmically, these *clusters* can then be described qualitatively (Tbl. 1) and presented visually (Fig. 1), creating products that have been used with great success in fields ranging from direct marketing (Evans, 1998), retail location selection (González-Benito & González-Benito, 2005), political campaigning (Robbin, 1980; Webber, 2006), and military recruitment (DeReu & Robbin, 1981), to social service resource allocation (Longley, 2005) : whether that service be in the field of health (Farr et al., 2008), education (Singleton & Longley, 2009), or policing (Ashby & Longley, 2005).

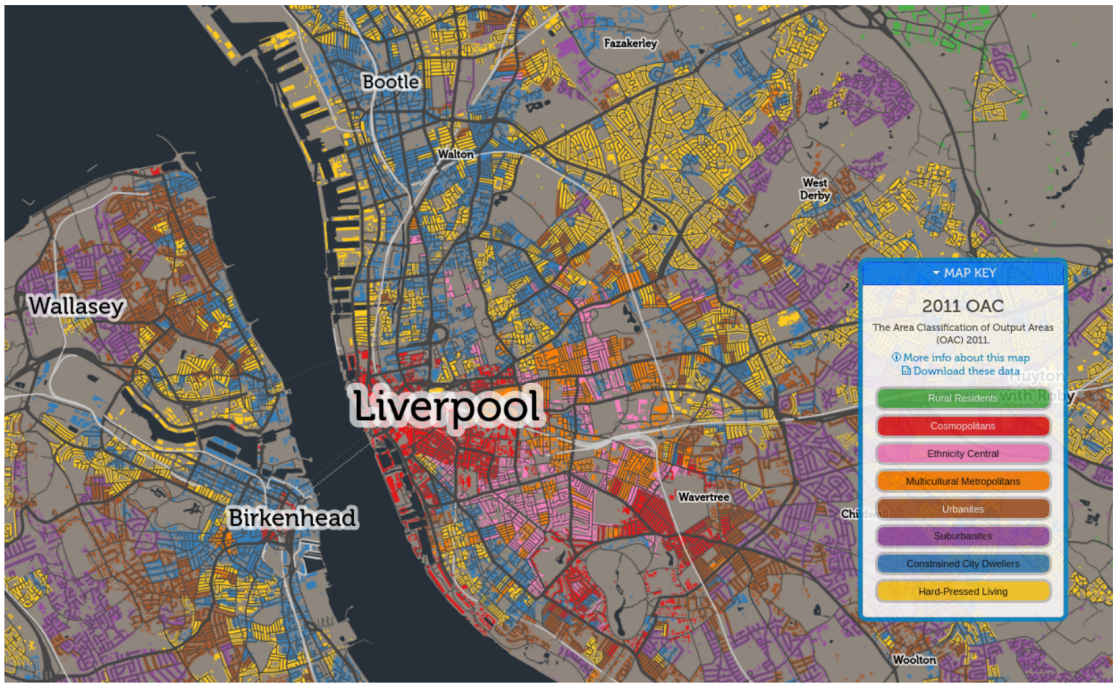

Figure 1: *An Example of Geodemographic Visualization: Choropleth Map showing 2011 Open Area Classification, after CDRC (2021)*

However, although it has been seen widespread adoption as a technique, there remain unresolved questions concerning its primary object, the neighbourhood unit (Petrović et al., 2020). Rather than reflecting a meaningful theoretically-grounded understanding of neighbourhood on-

tology, the units of geodemographic analysis are too often defined merely by data availability; or at a deeper level, by the administrative pragmatism and historical contigency responsible for defining the boundaries of data collection units.

To give credit where it is due, work such as that of Martin et al. (2001), which used a microsimulation approach to design census output zones maximizing social homogeneity, demonstrates that administrative data collection units do sometimes reflect a high level of sophistication. But perhaps here the problem is in fact presented even more starkly, for from a *statistical* point of view, it makes sense to design neighbourhood zones in such a way as to maximize social homogeneity. From a *social* point of view however, enforcing homogeneity is essentially *segregation*, and while as scholars we may be content merely to understand its causes (e.g. Schelling, 1969) rather than to resist it more directly (e.g. King, 1968), it becomes impossible to say much about whether and why neighbourhoods are (not) diverse, if our only way of defining them is in such terms.

Table 1: *An Example of Geodemographic Cluster Description: Groups and Subgroups from the 2011 OAC, created by Gale et al. (2016)*

Supergroup	Group
Rural residents	Farming communities
	Rural tenants
	Ageing rural dwellers
Cosmopolitans	Students around campus
	Inner city students
	Comfortable cosmopolitan
	Aspiring and affluent
Ethnicity central	Ethnic family life
	Endeavouring ethnic mix
	Ethnic dynamics
	Aspirational techies
Multicultural metropolitans	Rented family living
	Challenged Asian terraces
	Asian traits
Urbanites	Urban professionals and families
	Ageing urban living
Suburbanites	Suburban achievers
	Semi-detached suburbia
Constrained city dwellers	Challenged diversity
	Constrained flat dwellers
	White communities
	Ageing city dwellers

Supergroup	Group
	—
	Industrious communities
Hard-pressed living	Challenged terraced workers
	Hard-pressed ageing workers
	Migration and churn

As well as this ontological critique, geodemographic analysis has also been the subject of ethical critique, in which it has been portrayed as a prime example (Goss, 1995; Dalton & Thatcher, 2015) of what has since been described as *surveillance capitalism* (Zuboff, 2015, 2019). But while this may apply to proprietary systems, the criticism cannot be reasonably applied to open-source geodemographic analysis – indeed, to whatever degree one might be worried about the possibility of encroaching corporate *geosurveillance*, developing free and open alternatives to proprietary products is arguably a necessary strategy of resistance (Swanlund & Schuurman, 2019).

The aim of this paper is therefore to establish a more robust foundation for geodemographic analysis, by attempting to identify a theoretically rigorous unit of neighbourhood analysis, thus addressing the ontological critique; and to do so using open data and open-source software, thus resisting the ethical critique.

2. Literature Review

2.1. 'Geodemographics': Development and Debate

The word 'geodemographics' was coined by Jonathan Robbin (1980) to describe the marketing tool his company had developed (Webber & Burrows, 2018, p. 94). By classifying American residential zip codes into groups with similar demographic characteristics, and then giving each group a memorable label and summary description, he had created a product designed to simplify the process of targeting prospective customers and selecting promising retail locations (DeReu & Robbin, 1981). Robbin's tool combined the latest in marketing theory with cutting-edge methods in quantitative geographic sociology.

Market segmentation (Smith, 1956) solved the problem of a complex market of heterogenous customers by dividing it into several sub-markets of homogenous customers. *Demographics* – that is, population attributes such as age, sex, income, and ethnicity – offer a straightforward way of applying this strategy. The idea of *psychographics* (Wells, 1975) is then to understand the psychology of a typical customer from a given market segment, so as to anticipate their needs, desires, and trigger points. Robbin added these techniques to the *social area analysis* (Shevky & Bell, 1955) he had been applying as a doctoral candidate at New York University (Ricercar, 2021), assisting Edgar Borgatta in researching ways of classifying the social characteristics of American cities (Hadden & Borgatta, 1965).

Shevky & Bell (1955) offered a method of classifying a census tract by reducing the attributes of its census data into a simplied expression of just three factors, which they suggested both retained the important details of the data, and corresponded to the essential nature of contemporary society. Each datapoint could then be visualized by a small circle in a two-dimensional scatter-plot, with

the attributes along the x and y axes corresponding to the two more significant factors, and the circle's colour the third. The datapoints can then be divided up according to their position, and since their positional proximity is a function of their statistical similarity, census tracts with datapoints in the same segment can be considered as being of the same *type*.

Tryon (1968) achieved a more sophisticated way of grouping census tract datapoints, showing how his *cluster analysis* (Tryon, 1939) could identify 'clusters' of similar datapoints, avoiding the arbitrariness of simply segmenting the attribute space by intervals. Instead these could be detected by an iterative computational algorithm, made available in reproducible FORTRAN code (Tryon & Bailey, 1966). From a contemporary perspective, cluster analysis is a primary example of unsupervised machine learning (Géron, 2019), thus making geodemographic analysis one of the first, and arguably even the original, example of its application.

While clustering census tracts provides a way of segmenting them into similar groups, the *ecological fallacy* (Robinson, 1950) means it does not follow that individuals within tracts with similar population demographics are necessarily similar at the individual level. The argument for the likely homogeneity of census tract units is made by reference to the account of the *neighbourhood* as a *natural area* given by Park (1925), who argued that the twin forces of homophily and social influence will tend to segregate an urban population into a "mosaic of social worlds" (Wirth, 1938). Park was the central figure of the *Chicago School* (Abbott, 2017), whose distinctive *human ecology* combined the empirical investigation exemplified by Booth (1904) with the more theoretical sociology of Simmel (1908), with whom he had studied in Germany.

This account of geodemographic historiography is expressed in detail in the collaborative monographs of Harris et al. (2005) and Webber & Burrows (2018). Booth's urban poverty maps offer "the first example of applied geodemographics" (Harris et al., 2005, p. 30), the Chicago School's theory of urban *natural areas* (Fig. 2) provides the necessary "conceptual definition…[for] neighbourhood analysis" [p.39], the increasing availability of census data then stimulates the development of quantitative social area analysis and subsequent factorial ecology [pp.39-40], paving the way for the emergence of geodemographic products [p.55].

While Robbin coined the term 'geodemographics' and successfully turned it into a profitable commercial product, his impact on the academic understanding of neighbourhoods was quite limited. He dropped out of his PhD without completing it, after his supervisor left to take up a post at another university (Ricercar, 2021), and it has subsequently been the case in North America that, as a method for studying neighbourhood dynamics and effects, "many academic social scientists ignore geodemographics" (Reibel, 2011, p. 310).

In Britain however, the situation is quite different, largely because of the different circumstances of its separate development by Richard Webber. Independently from Robbin, Webber was also applying cluster analysis to census data; first, for the purpose of helping the Liverpool City Council identify priority areas for social service provision (1975), and then on a national scale (1978). Although Webber did go on to work in the marketing industry, the initial context of his research for a public authority meant that his work was published openly, rather than being a proprietary secret. There has then been a continued tradition of free, national, *open* (Singleton et al., 2016) geodemographic typologies of Britain, produced using the data from the 1981 (Charlton et al., 1985), 1991 (Blake & Openshaw, 1994), 2001 (Vickers & Rees, 2007), and 2011 (Gale, 2014; Gale et al., 2016) censuses.

This has allowed robust academic debate about the validity and utility of geodemographic

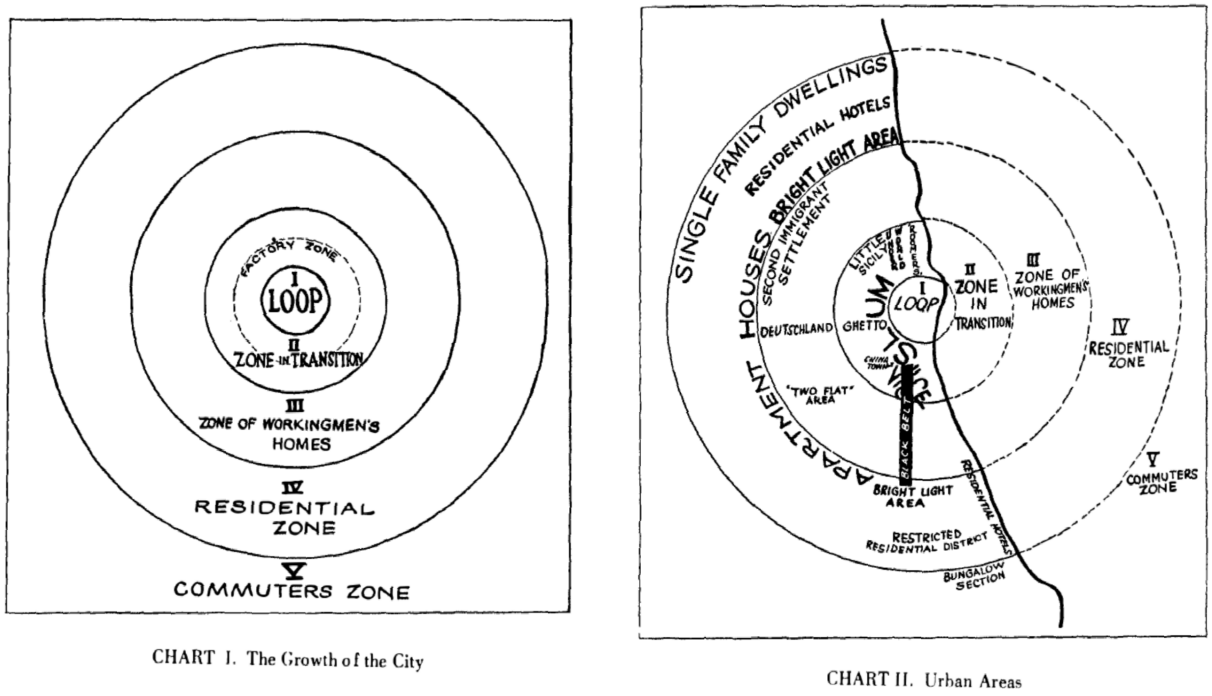

CHART I. The Growth of the City

CHART II. Urban Areas

Figure 2: *The Chicago Concentric Zone Diagram, after Burgess (1925)*

typologies. Openshaw & Gillard (1978) showed the instability of clustered classifications by demonstrating their sensitivity to subjective decisions at multiple points in their construction, including "the selection of variables, the choice of algorithms and methods, and various data management operations" (p.101), and concluded that such classifications in general, and Webber's 1975 study in particular, "should not be used until they can be replicated at the individual level" (p.118). Openshaw et al. (1980) repeated the warning "for all users to be aware of the practical limitations of [Webber's subsequent national] Classifications… [as] they are unsuitable or many of the applications that have been suggested" (p.438). Webber apparently was apparently unaware of the first critique, but the second quickly provoked a thorough rebuttal, in which Webber (1980) concluded that his critics were not "in touch with either public policy or the commercial world" (p.449). Presumably, the point was well taken, as Openshaw (1985) then applied census-data cluster classification to rural areas, complaining of the resistance of government departments to such methods, and of their "preference for old fashioned pre-computer age techniques" (p.286).

Openshaw (1997) then found himself in another notable controversy concerning geodemographics, this time offering a defense, when he offered a widely-ranging response to the various criticisms of Geographic Information Systems gathered by Pickles (1995). The point which stands out as having the most continued validity in his discussion is his suggestion that "data protection legislation" can mitigate some of the dangers of unfettered technological surveillance, foreseeing the need for the sort of data protection legislation now established by statutes such as GDPR (2016).

In our contemporary context, Gilbert (2021) offers another helpful response to the suggestion that gathering and profiting from people's data is a sort of exploitative 'data colonialism' (Thatcher et al., 2016), in which an intrinsically valuable resource is being unfairly extracted. Rather than being 'the new oil' (e.g. Szczepański, 2020), Gilbert, whose describes himself as a 'data optimist,' suggests that a better metaphor for big data would in fact be 'the new manure':

"a mundane by-product of life" [p.36], which, like manure that is processed into fertiliser, only has economic value because there are businesses that have invested in processing it into something useful.

We cannot in this paper attempt a comprehensive evaluation of Gilbert's attempt to defend big data analytics and develop a positive account of *digital legitimacy* (Greene & Gilbert, 2020). For our present purposes it will suffice to note that on the one hand, a positive argument can be made for it, while on the other hand, to whatever extent the negative assessment is considered valid, the development of free and open alternatives would seem to be a necessary strategy of resistance (Swanlund & Schuurman, 2019). But regarding the ontological question of how to define a neighbourhood, more must be said.

2.2. Defining Neighbourhoods: Problems and Possibilities

One possible response to the question would be to suggest that in fact the problem of neighbourhood definition is nothing more than a particular instance of the the more general *Modifiable Areal Unit Problem* (MAUP), described with typical clarity by Openshaw (1983), but in fact identified fifty years previously by Gehlke & Biehl (1934). The problem is a profound one for quantitative analysis involving spatial data, for it observes that the same basic dataset can yield quite different statistical results depending on the specific ways that its data has been aggregated. The effect is found not only when data is gathered at different scales, but even when it is aggregated at the same scale with differing boundaries. A contemporary example is shown in Fig. 3, in which Singleton & Cheshire (2021) demonstrate how, depending on the size of the population of the areal unit used for analysis, COVID-19 rates can appear "as low as 295 per 100,000 people or as high as 736 per 100,000."

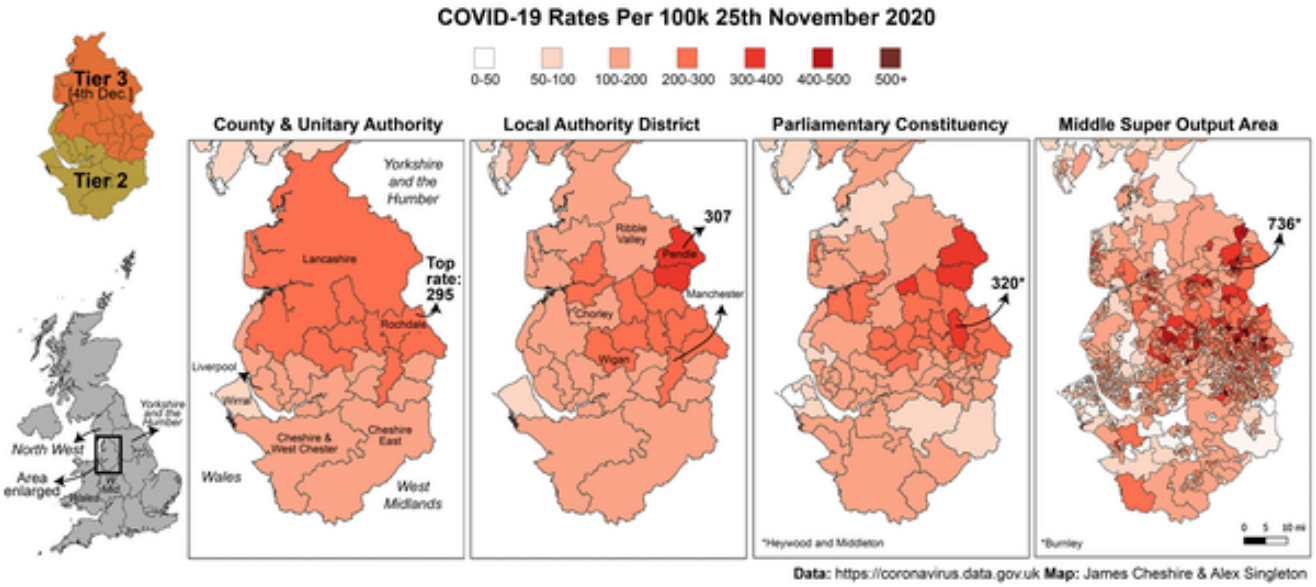

Figure 3: *Demonstrating the Modifiable Areal Unit Problem with COVID-19 Rates, after Singleton & Cheshire (2021)*

A slightly different challenge to quantitative spatial analysis is the *Uncertain Geographic Context Problem* (UGCoP) described by Kwan (2012). The MAUP is a problem of how to

aggregate individual datapoints into collective units. It observes that different ways of aggregating spatially located datapoints into broader areal units may lead to different statistical conclusions, and therefore warns against treating any particular areal unit as authoritative. The UGCoP, however, points out that even when only considering a single social datapoint, that point is associated with a human individual who will have experienced exposure to relevant contextual influences in a variety of geographical contexts not limited to the point at which they live, or perhaps were interviewed, the details of which will in general be unknown to the researcher. But rather than merely confounding the issue further, Kwan frames the problem in such a way as to suggest that there is in fact some "true causally relevant geographic context" (p.959).

For Kwan, this suggests a turn "from location to movement, from place to mobility, and from space to space-time" (p.966), and she suggests "using GPS data to delineate activity spaces" (p.965). This is all very well, but in heeding her suggestion we might perhaps stray too far from our own topic of geodemographic ontology. But nevertheless, the suggestion that for every spatial effect, there must be *some* true causally relevant geographic context, rekindles the hope that even in considering a general typology of residential neighbourhoods, such a context might be found.

What then are the causally relevant contextual elements that make a neighbourhood a neighbourhood? We find ourselves returning to the observation made by Park (1925, p. 1), that "there are forces at work… within the limits of any natural area of human habitation… which tend to bring about an orderly and typical grouping of its population and institutions." Does Park's claim still hold? And if so, with which "forces" should we be primarily concerned? Is it "the economic organization of the city… based on the division of labour" (p.2)? Or rather "racial, cultural and vocational interests" (p.11)? Or "the breaking down of local attachments and the weakening of the restraints and inhibitions of the primary [family] group, under the influence of the urban environment" (p.25)? Or it is the economic expansion of the city, and the accompanying "tendency of each inner zone to extend its area by the invasion of the next outer zone" (Burgess, 1925, p. 51), Fig. 2?

An impressive case is made by Galster (2019) that "to understand the causes and effects of neighborhoods one must embed them in a framework in which four spatial levels–metropolitan, local jurisdiction, neighbourhood, and individual–are interconnected in mutually causal ways" (Fig. 4). At the individual level, we have the mobility and investment behaviours of individual households, dwelling owners and developers. These both influence and are influenced by the demographic, social, economic and physical characteristics of the household's surrounding neighbourhood. Simultaneously, these neighbourhood characteristics are engaged in circular interaction with public and private service providers operating at a broader geographic scale. And all of this occurs within the context of the regional housing market.

Galster credits his multilevel model to the inspiration of Suttles (1972), whose "groundbreaking observation that people are cognitive of four distinct spatial levels of neighbourhood" (p.39). At Suttles' time of writing, the 'natural area' concept of Park, Burgess, and the interwar Chicago School, had fallen thoroughly out of fashion. Alihan (1938) had concluded that although "the ecological school [was] one of the most definite and influential schools in American sociology" (p.xi), "the concept 'natural area,' so fundamental to human ecology, has not as yet been consistently defined and logically classified… [and] no amount of empirical investigation can rectify the inconsistencies inherent in the theoretical statements pertaining to it" (p.240). Refusing to accept Alihan's damning verdict, Suttles (1972, p. 21) attempted "to resurrect the concept… and show that it may still be usefully applied to urban areas," noting the need to consider both the

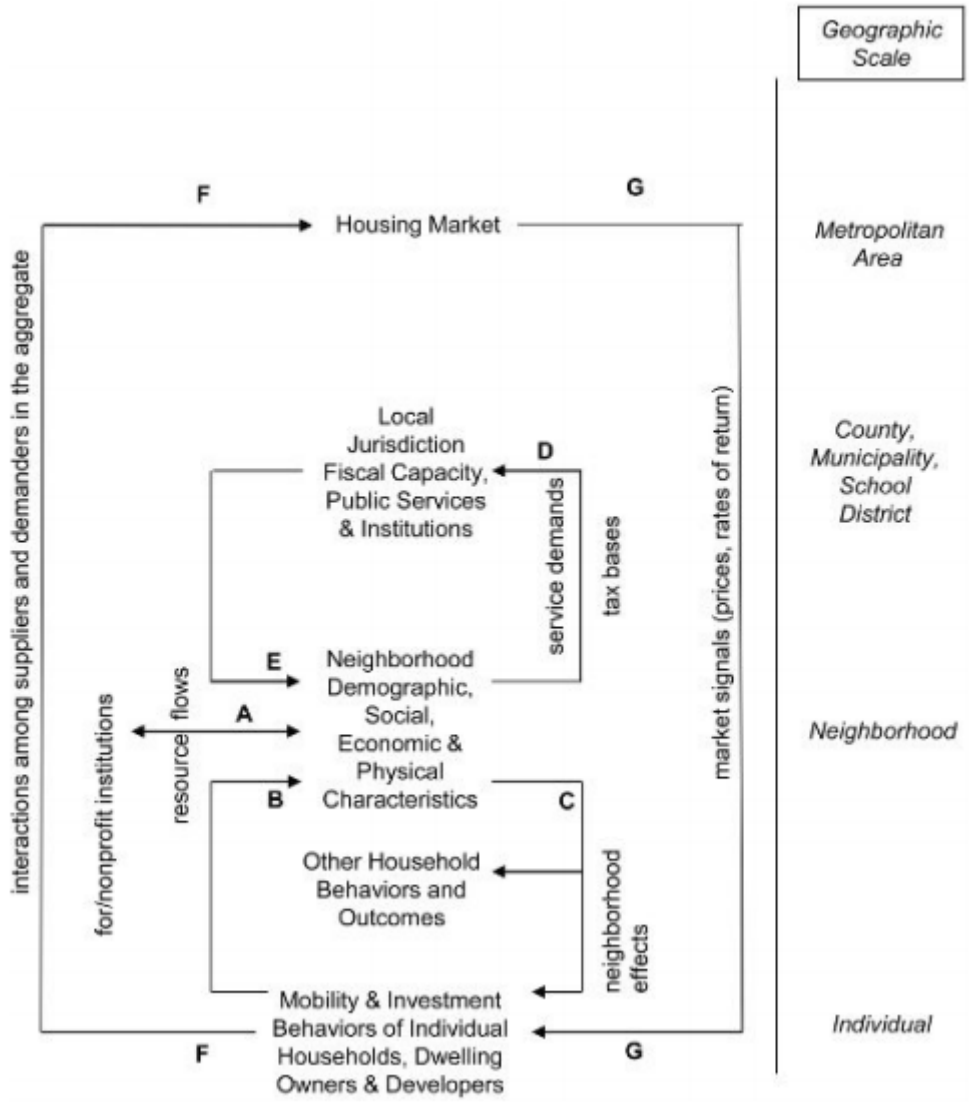

Figure 4: *Holistic, multilevel, circular causation model of neighbourhoods, after Galster (2019)*

"physical structure of the city," and "the cognitive map which residents have."

Drawing on the suggestion that there are analogies between human social behaviour and the ideas of animal *territoriality* developed by zoologists such as Lipitz (1969) and Morris (1967), Suttles (1972) suggests that neighbourhood community "is best conceived of as a pyramid of progressively more inclusive groupings" (p.45), and identifies four relevant levels of analysis: the 'face-block,' the 'defended neighbourhood,' the 'community of limited liability,' and the 'expanded community of limited liability.' At the smallest level, the *face-block* "is the smallest discrete areal unit other than the household which [residents] can point to" (p.56). Suttles takes for granted that his readers will understand what a 'face-block' is, but confuses matters somewhat by introducing the concept together with the loose local network of acquaintances selected "because they are known from shared conditions of residence" (p.55). In an otherwise excellent review of the literature, Chaskin (1997) incorrectly identifies Suttles' definitions of 'local network' and 'face-block,' and suggests that a face-block has no precise residential identification. But in fact Suttles notes that unlike the loose network which is "unlikely to have any sharp boundaries" (p.55), the face-block is notable specifically for having an areal basis so clear that parents are able to use it for instructing their children (p.56). For an explicitly articulated definition, we must turn to Grannis (2009, p. 31), who explains (consistently with Suttles' usage) that "the face block includes all of the dwellings that front on the same street and are situated between the first cross streets, of any type, encountered in both directions away from the respondent's house."

The face-block is of particular interest, because it offers a unit of analysis that is primary from both perspectives necessary to a robust neighbourhood ontology, both those of physical structure and of cognitive mapping. In the last decade the explosion of ubiquitous urban data (Arribas-Bel, 2014) has catalyzed significant advances in the morphological analysis of urban physical structure, with the studies of Barthelemy (2017), Louf & Barthelemy (2014), Schirmer & Axhausen (2016), Boeing (2019) (2020a), and Fleischmann, Romice, et al. (2020) of particular note. But with regard to the latter point of cognitive social maps, although the essential ideas have been well-established since the studies of Gould & White (1974) and Lynch (1960), there remains more work to be done in integrating these concepts into large-scale analyses that take advantage of the detailed data now available. The attempt of Lai et al. (2020) to profile urban places based on geotagged Twitter data for London suggests one possible direction of enquiry. But if we can show more generally that there are strong theoretical reasons for the significance of the face-block, then we can use the analytic tools already developed for morphological analysis and claim them for more social investigation as well.

At a larger level, it is well-established that the structural features of major roads, railway lines and rivers – Jacobs (1961) refers to them as *border vacuums* – are also perceived as social boundaries. Burgess (1925) acknowledges that his simplified theory of urban economic expansion is complicated "by the lake front, the Chicago River, railroad lines, [and] historical factors in the location of industry" (p.52), and these complications are shown on Chart II of his well-known Concentric Zone Diagram (Fig. 2). But Grannis (2009) demonstrates that not only are neighbourhoods defined by the way that urban areas are *divided* by major roads (and railways, rivers, etc.), but that for the households within the same set of boundaries to be accessible to each other, they also need to be *connected* by safe, walkable pedestrian streets – that is, by contiguous residential face-blocks.

Grannis roots his argument in a simple account of how neighbouring relationships necessarily develop along a natural scale of relational availability (Tbl. 2). At the lowest level (0), we have the

situation where there is simply no availability at all – and thus there is no neighbourly relationship. The most basic level (1) in actually being neighbours is geographical availability, for "proximity is essential to the very definition of neighbouring" (p.19). The next level (2) is achieved when passive contact takes place, as neighbours "unintentionally encounter each other on a regular basis." The relationship can then develop to involve intentional contact (level 3) and mutual trust (level 4).

Table 2: *Levels of Relational Availability, (after Grannis, 2009)*

Level	Relational Availability
0	No availability
1	Geographic availability
2	Passive contact
3	Intentional contact
4	Mutual trust

He then suggests that these individual neighbourly relations concatenate to form networks corresponding to the relevant relational stages (pp.37-47). In particular, what becomes apparent is that to "transcend the network of geographic availability … is logically impossible" (p.40). He thus concludes that "the maximal concatenation of contiguous face blocks… represents the maximal consolidation of individual residents' potential contact with each other" (p.42).

We can supplement Grannis' theory of how contiguous walkable face-block networks necessarily bound neighbourhood networks, with some of the insights of Jacobs (1961) about how a city's streets need to serve the vital social purposes of creating a natural place for public contact (pp.72-96), and of providing the "eyes upon the streets" (p.45) necessary to induce the social restraint which makes for public safety (pp.37-71).

3. Methodology

3.1. Basic Idea: Defining Neighbourhood Units with Street Network Geometry

The essential idea of this paper is to define neighbourhood units for the whole of Britain based on the street network, rather than the administrative boundaries generally used by geodemographic typologies. Initially I had thought that it would be sufficient to perform a tessellation based on treating roads (and railways and rivers) as borders which naturally divide neighbourhood areas (Fig. 5, left).

This is the approach taken by the Urban Grammar project of Fleischmann & Arribas-Bel (2021), which first generates tessellated 'enclosures' and then subdivides them into Voronoi cells based on building footprint polygons, as described by Fleischmann, Feliciotti, et al. (2020). It was very helpful to learn from the implementation of this method, shared openly on GitHub (Dabbish et al., 2012) in reproducible Jupyter notebooks (Kluyver et al., 2016; Randles et al., 2017; Boeing & Arribas-Bel, 2021).

However, in reviewing the literature, it became apparent that while certainly motorways and railways act as dividing boundaries of social space, residential streets perform the opposite role,

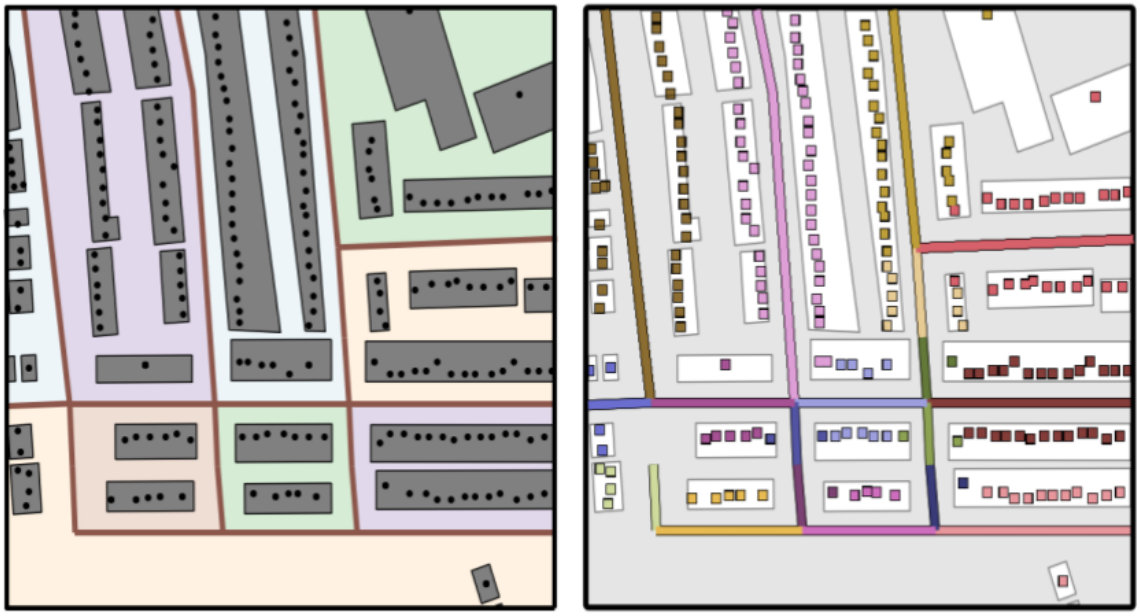

Figure 5: *Visual Comparison between (left) Tessellated 'Blocks' divided by bordering streets, and (right) natural Face-Blocks connected by nearest streets*

connecting rather than dividing. My goal was therefore to operationalize a way of algorithmically describing the face-block (Fig. 5, right) that Suttles (1972) and Grannis (2009) have demonstrated is a fundamental building-block of neighbourhood space, and of identifying the connected community networks formed of concatenated residential face-blocks.

3.2. Computational Setup: Open Data and Free Open-Source Software

To do this I used data from Ordnance Survey's OpenData product suite (Tbl. 3), which was launched in 2010 (Lilley, 2011), but has had to navigate a certain tensions between various economic, political, and legal interests (Field, 2010). In particular, the Ordnance Survey is obliged by statute to make its data widely available, but also to finance its operations independently through a commercial licensing model (Birss, 2019), a business model which Boswarva (2019) has suggested "combines the worst features of state monopoly and rentier capitalism." One significant omission in the catalogue of British open data is an authoritative national address dataset (Wells, 2021).

Table 3: *Ordnance Survey Open Data*

ID	Version	Download Size	Useful Contents
OpenRoads	2021-04	900M	Road segments and intersecting nodes
OpenUPRN	2021-08	2.0G	Properties
OpenMapLocal	2021-04	3.6G	Building polygons and railway lines
OpenRivers	2021-04	52M	Rivers
Strategi	2016-01	39M	Simplified coastline

In 2018, the British government's commitment to open geospatial data was reiterated with the creation of the GeoSpatial Commission (Office}, 2020). While this has not led to the release of authoritative address data, it has at least led to the release of the Unique Property Reference Number (UPRN) dataset, which includes a geographic location and unique numeric identifier for every addressable location in Great Britain found in the premium OS AddressBase products. Addressable properties given a UPRN include objects such as bus shelters, lamp posts, and public toilets (Greenwood & Brandwood, 2020), so although we know that neighbourhood household properties form a subset of the UPRN data, we need to somehow eliminate those properties that we are not interested in.

This was done by using the 'Buildings' polygons from the OpenMapLocal dataset. On their own these cannot either provide us with a proxy for neighbourhood households, for they do not give the boundaries of individual property plots – although Fleischmann & Arribas-Bel (2021) seem to use them as if they do – but only of disconnected architectural structures. So for example a terraced row of houses is shown as a single building. But by combining the two sets of data, it becomes possible to exclude property reference points which refer to outside objects.

For street (or road) network data, there were several possible candidates, including the USRN dataset (which, like the UPRN, offers unique locational reference numbers, but for streets); the 'Roads' layer of the OpenMapLocal dataset, and the OpenRoads dataset (Tbl. 4). I used the last of these three options.

Table 4: *Summary Statistics for OS OpenRoads Street Segments*

Road Function	Count (%)	Length/m (Q1)	Length/m (Q2)	Length/m (Q3)	IQR
Motorway	7,124 (0.1)	85	327	570	485
A Road	285,536 (7.5)	33	69	169	136
B Road	159,415 (4.2)	39	85	204	165
Minor Road	676,167 (17.9)	43	101	275	232
Local Road	1,650,781 (43.8)	43	68	109	66
Local Access Road	45,307 (1.2)	44	73	137	93
Restricted Local Access Road	803,087 (21.3)	50	100	200	150
Secondary Access Road	133,522 (3.5)	39	53	86	47
Total	3,760,939 (100.0)	43	76	146	103

I also used the simplified coastline data from the Strategi product. For our purposes there is no need for the other more detailed (and therefore more computationally resource-consuming) water-line boundaries also available. Together with the 'Railways' layer of the OpenMapLocal dataset, the OpenRivers dataset, and the major roads from the OpenRoads dataset, these provided a full set of boundaries for tessellating Britain into naturally bounded areas.

To simplify setup, and to make it easy not only to make my analysis reproducible across different machines, but to make it easy to restore my computational environment if and when necessary, I used Docker, which has become accepted as a powerful solution for reproducible research and collaborative software development (Boettiger, 2015). Docker allows the required configuration to be specified as code, run in an isolated container, and reproduced straightforwardly simply by

building an image from the relevant Dockerfile.

I made use of the Geographic Data Science notebook stack maintained by Arribas-Bel (2019), which extends the official Jupyter Docker Stack with a comprehensive set of geospatial Python libraries (of which Fleischmann et al., 2021 give a full description). I coupled this with a separate Docker container running the most up-to-date version of PostGIS, which extends the excellent open-source database PostgreSQL (Momjian, 2001) with the functionality to make spatial queries.

Fleischmann & Arribas-Bel (2021) download the datasets from the Ordnance Survey API within a Jupyter notebook, and save them to a PostGIS database by first loading them into a Geopandas dataframe, but while this method is feasible with the smaller OS datasets, it became incredibly slow when dealing with almost 40 million rows of UPRN data. When I instead used the GDAL ogr2ogr tool (Contributors}, 2021) – again, just by pulling the most recent official Docker image – to inject the data from the downloaded GeoPackage directly into the database, the process became much quicker: shortening from over an hour, to just a few minutes.

3.3. Conceptual Definition: Metric Spaces, Topological Neighbourhoods, and Walkable (Hyper)Graphs

In order to analyze the propinquity of neighbours, we need to understand how to calculate the proximity of points.

Given some set X, a *metric* $d : X^2 \to \mathbb{R}$ is a function with the following properties:

- $d(x, y) \geq 0$ for all x, $y \in X$.
- $d(x, y) = 0$ if and only if $x = y$.
- $d(x, y) = d(y, x)$ for all x, $y \in X$.
- $d(x, y) + d(y, z) \geq d(x, z)$ for all x, y, $z \in X$.

Perhaps the most familiar metric is the Euclidean norm $\|\|_2$ on $\mathbb{R}^2$, derived from the Pythagorean theorem and easily extended to $\mathbb{R}^n$:

$$\|\mathbf{x}\|_2 = \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2}$$

This is the metric that our analysis uses to calculate the distance between a property and its nearest street or building.

It has become a commonplace in certain streams of critical geography to criticize quantitive geography for being "intrinsically tied to absolute, Euclidean and Cartesian perspectives on space" (O'Sullivan et al., 2018). A pedantically quantitative geographer might reply that actually the Earth is not a flat plane on which the Euclidean metric suffices, but rather (very nearly) an oblate spheroid (Mathews & Shapiro, 1992). To calculate the distance between two points on the surface of a sphere, we should calculate the *great-circle distance*, that is the distance along the shorter arc of the circle which cuts through both points and the centre of the sphere. For a spheroid, the formulae of Vincenty (1975a), (1975b) provide iterative methods for calculating the distance precisely.

In fact, since the geographical positions of our data are given by reference to the British National Grid, one of several coordinate reference systems commonly used to describe location in Britain (Tbl. 5), and since the distances we are here interested in are all small, we will ignore the earth's curvature and use the Euclidean metric to calculate geometric distances.

Table 5: *Some Reference Systems commonly used for describing location in Britain*

Name	EPSG	Used by	Coordinates
OSGB 1936 / British National Grid	27700	Ordnance Survey	Easting, Northing
World Geodetic System 1984	4326	GPS navigation	Latitude, Longitude
European Terrestrial Reference System 1989	4258	European Union	Latitude, Longitude
Pseudo-Mercator (Spherical)	3857	Google Maps	Easting, Northing

But the critical geographers are right to assert that humans inhabit a relational and social space which should not be blindly assumed to have the features of a mathematical metric space. In particular, the requirement that a metric is *symmetric* – that is $d(x, y) = d(y, x)$ for all x, $y \in X$ – seems too strong to hold in many situations. A simple example of where it does not hold, would be that of distances by road where the road network includes one way streets. Or if we wanted we could allude to the asymmetries in some Foucauldian web of knowledge-power (Klauser, 2013)

The mathematical study of structured spaces more general than metric spaces is called *topology* (Korner, 2013). Specifically, if we have a set X, then τ the set of subsets of X is a topology if the following hold:

- The empty set $\emptyset \in \tau$ and the space $X \in \tau$.
- If $U_\alpha \in \tau$ for all $\alpha \in A$, then $\bigcup_{\alpha \in A} U_\alpha \in \tau$.
- If $U_j \in \tau$ for all $1 \leq j \leq n$, then $\bigcap_{j=1}^{n} U_j \in \tau$.

It seems relevant to note that mathematical topology includes a well-defined concept of *neighbourhood*. Given a topological space (X, τ), then for $x \in X$, we say that N is a neighbourhood of x if we can find $U \in \tau$ with $x \in U \subseteq N$.

This does not answer the question of how to define a socially meaningful *human* neighbourhood, but it is interesting that even from the precise perspective of pure mathematics, there is not necessarily a unique answer to the question of what the neighbourhood is of a given element in a set; rather, there may be multiple containing sets that fulfil the conditions necessary to be a neighbourhood.

Now, regardless of whether it be Euclidean, spherical, or oblate spheroid, rather than considering the whole continuous space of the plane of the British National Grid, all we really want to consider are the discrete points that represent households, and the (continuous) lines that represent streets. The Ordnance Survey OpenRoads dataset models the British road network as a graph of undirected street segment *edges* connecting the *nodes* of street intersections.

In mathematical graph theory, a *graph* is an ordered tuple $G = (V, E)$, consisting of a set of *nodes* (or *vertices*) $V = \{v_i\}$, and a set of *edges* $E = \{e_{ij}\}$, where the edge $e_{i,j}$ is the ordered pair (i, j) representing some connection from the *source* v_i to the *target* v_j. A *weighted* graph has a

function $w : E \to \mathbb{R}$; in our example of streets, a possible weight function would be the length of the street. In an *undirected* graph, the presence of the edge $e_{i,j}$ implies and is implied by the reverse edge $e_{i,j}$, and there are no one-way connections; in a *directed* graph, this is not true.

Given a graph G, we can describe a *walk* of length L as a sequence of adjacent (but not necessarily distinct) nodes $(v_0, ..., v_L)$; or, equivalently, as a sequence of edges $e_{0,1}, ..., e_{L-1,L}$. If we have an undirected graph with positive weights, then we could define a metric on the graph by the shortest walk between any two points.

The idea of a graph, in which edges connect precisely two points, can be extended to that of a *hypergraph*, where *hyperedges* can connect any number of points. For our purposes, if we consider a street segment to connect to each other all the properties which have that street as their nearest street, and also to the nodes at its ends, then our graph of street segment edges becomes a hypergraph in which each property connects to exactly one single edge.

A *face-block F* is then the set of properties connected by any street segment edge, and the power-set of all face-blocks forms a topology, in which the (topological) neighbourhood of any property is any set of properties which includes its face-block. Now, if we were to return to the Cartesian plane with the Euclidean metric, a tessellation of the plane also forms a topology, and any boundary that encloses an entire tessellated tile is a (topological) neighbourhood of any point within that tile. Which illustrates what we have already said, that a precise definition of 'neighbourhood' does not imply that there can only be one way of bounding a validly defined neighbourhood for a given element.

4. Data Analysis

Having obtained the necessary data, set up our computational environment, and understood the relevant concepts, we are now ready to do some data analysis. Unlike much statistical analysis, my approach here is not one of seeking to confirm some particular hypothesis (or to reject some null hypothesis). As Tukey (1977, p. 3) says, "unless exploratory data analysis uncovers indications… there is likely to be nothing for confirmatory data analysis to consider." But my aim has been also to do this exploration of the data in such a way as to create reusable tools (cf. Boeing, 2020b) that myself and others could continue to use. The code written for the purpose of this analysis is therefore available on Github as the installable Python module `nbhd`.

4.1. Neighbourhoods: Residential Face-Blocks and Connected Street Networks

As we have already seen, a *face-block* is the set of households which front on to the same street segment. We have street segment data, but we do not have data specifically about household addresses. We do however have geographically located property reference points, and building polygons, and we assume that a household property must exist within a building. The building polygon data is primarily cartographic, and so not intended to be precisely accurate. In particular, Survey (2019, p. 10) explain that the OpenMapLocal dataset has been subject to the *simplification*, *exaggeration*, and *aggregation* typical of cartographic map generalization. So, for example a terraced block of houses appears as a single building.

But by combining the building polygon data with the property reference locations, we have a proxy for how many houses make up that terrace, and by dividing the footprint of the building

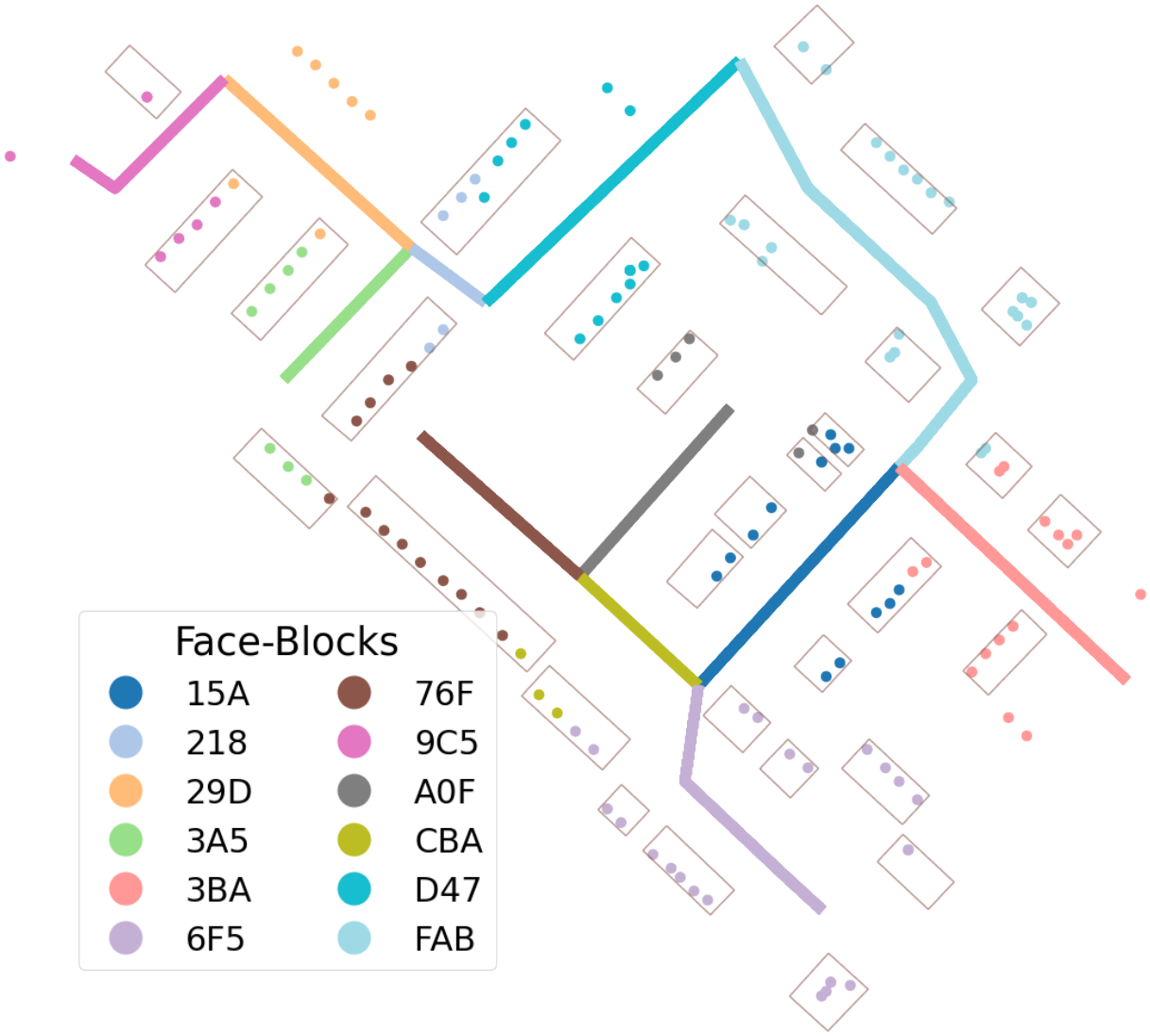

Figure 6: *Face-Blocks in a Connected Street Neighbourhood*

polygon by the number of property reference points contained by the building's geometry, we have a proxy for how big those houses are. If the average footprint per property of a particular building is over a certain threshold (I chose 250m^2 as a safe default, but it is a customizable parameter within the relevant function), then it is assumed to be some institutional building (say, a school or a shopping mall), and disregarded from our analysis. Within the constraints of our current data however, we have no evident way of distinguishing between a household and a small shop. One justification for allowing this might be to invoke the *need for mixed primary uses* that Jacobs (1961) observed was a necessary feature of healthy urban neighbourhoods.

Table 6: *Selected Statistics for Every Face-Block in a Connected Street Neighbourhood*

ID	Road Name	Properties	Buildings	Neighbours	Length/m	Footprint/m ^{2}
76F	Cameo Close	13	3	2	46.8	688.6
CBA	Cameo Close	3	2	4	35.4	163.7
FAB	Conwy Drive	32	6	3	112.6	872.8
15A	Conwy Drive	14	6	4	65.1	526.5
6F5	Conwy Drive	25	8	2	62.7	816.9
D47	Jade Road	17	3	2	76.9	551.5
3BA	Montgomery Way	15	6	2	68.6	451.1
9C5	Opal Close	7	3	1	45.2	316.5
3A5	Pearl Way	7	2	2	40.3	349.3
29D	White Rock Street	7	3	3	55.2	369.9
A0F	White Rock Street	6	3	2	49.6	186.4
218	White Rock Street	5	2	3	20.3	252.1

Given that we do not have specific household address data, we certainly do not have data showing exactly which street segment the household's front door opens on to. So we simply assign each property reference point (within a building polygon) to its nearest street segment. Naively, we could import the data for our property points, and for our street segments, calculate the distances between each pair, and order them in ascending order; but if we were to do this in Python, even making use of the Pythonic geospatial stack, it is very slow even for a fairly small number of properties and street segments. Instead, our PostGIS database is able to do a *k nearest-neighbours search* very quickly with a query that combines a spatial distance condition with a *lateral join*, and takes this general form:

```
SELECT t1.id AS t1_id,
  t1.geometry AS t1_geometry,
  t2.id AS t2_id,
  t2.geometry AS t2_geometry,
  ST_Distance(t1.geometry, t2.geometry) AS dist
FROM table1 AS t1
    CROSS JOIN LATERAL (
      SELECT t2.*
```

```
    FROM table2 AS t2
    ORDER BY t1.geometry <-> t2.geometry
    LIMIT {k}
) AS t2 ;
```

Having assigned properties to their nearest street segment, we can calculate the mean length per property of the street segment. As with a building's footprint per property score, we also apply a threshold to a street segment's length per property. By default this is set at 50m, but again it is a variable parameter in the relevant function.

There are sometimes points in the OpenRoads dataset where from a vehicular perspective one road ends and does not connect to another, though the nodes at either end are so close as to be almost touching. I can say for certain that in at least one case this is a result the presence of bollards, which mean that for a vehicle there is no connected thoroughfare – but for the pedestrian walking around their own neighbourhood there would be. I therefore relabel nodes within a certain short radius (by default 5m) of each other with the same label, which means that cases like this are counted as connected roads.

Figure 7: *Neighbourhoods of Connected Face-Blocks*

Having done all this, I use Python's `networkx` library to create a network graph from the `pandas` dataframe containing the start and end nodes for each face-block's street segment, and find the connected subcomponents of the graph. Each can then be given a different label, and we can colour each different connected neighbourhood with a different colour and test by visual inspection whether our algorithm has performed as desired. Fig. 7 shows the method applied to the area of North Liverpool surrounding Newsham Park, which happens to be a neighbourhood of my own house. The streets and building polygons are coloured to indicate membership of the same connected neighbourhood. The black dots all represent properties located within the

geometry of building polygons, although I have not coloured networks of less than ten face-blocks, which is why some property points have no visible building polygon.

Broadly, we can see that the method works – contiguous areas seem to be connected, and the algorithm is sensitive to the subtleties of the street topology. For example the blue neighbourhood `01` and the orange neighbourhood `02` seem at first glance to be touching – surely they must form a single neighbourhood! And yet our method is able to correctly discern that actually there is no connecting thoroughfare that would make allow residents of the geometrically adjacent buildings to naturally bump into each other and so begin the haphazard journey up the scale of neighbourly relationship.

It might seem odd that the teal colour of neighbourhood `10` is highlighted as a distinct neighbourhood in spite of being tightly enclosed by the green of neighbourhood `3`. There appear to be two small connecting roads on the north and north-west vertices that have gone unnoticed. Presumably there is a bug in my implementation of the algorithm (for which I have now opened an issue on the `nbhd` package's Github repository. It does however provide a conveniently small connected neighbourhood that allows us to show (Fig. 6) and tell (Tbl. 6) more clearly exactly what we are able to say about a single connected network of face-blocks than we would be able to do with a larger group where in this context we lack the space to display all the details.

4.2. Nationwide Implementation: Embarrassingly Parallel Neighbourhoods

We now seek to apply our method to the whole of Britain. To do this we first divide up the area into what I mentally refer to as *pixels*: square boxes created by tessellating the area of the British National Grid with horizontal and vertical lines at intervals of 2000m. I then used the polygonized the Strategi coastline linestrings to create polygons for all the British isles, although for this paper the analysis has only been applied to mainland Britain. The pixels that did not intersect with mainland Britain were then discarded, leaving 56,849 pixels which did intersect.

The process of identifying face-blocks and connected street segment networks then turns out to be *embarrassingly parallel*, a phrase first coined by Moler (1986) for computational problems which experience dramatically effortless gains in processing speed when addressed as multiple small problems simultaneously, instead of sequentially as one very large problem. Such parallelization is often associated with large cloud computing clusters, and technologies such as Hadoop and Spark, but Python includes a `multiprocessing` module which allows one to take advantage of the same idea of the MapReduce method (Dean & Ghemawat, 2004) of mapping a function to the awaiting elements of data which need processing.

Table 7: *Summary Statistics of Connected Street Neighbourhoods*

	μ	σ	Q1	Median	Q3	IQR
Properties	12.6	8.1	7.0	12.6	14.2	7.2
Buildings	3.9	1.9	2.5	3.0	4.4	1.9
Neighbours	2.5	0.8	2.0	2.0	3.0	1.0
Length	56.6	21.7	42.7	55.2	63.3	20.5
Length/property	4.9	2.5	3.3	4.1	5.1	1.8
Total Footprint	462.1	216.6	284.3	451.1	532.8	248.4
Median Property Footprint	42.2	11.1	32.3	50.0	50.3	18.0

	μ	σ	Q1	Median	Q3	IQR
Mean Property Footprint	42.0	9.4	32.7	42.0	50.1	17.4

On my ten-core Intel Xeon desktop workstation, I was able to run the analysis for the whole of Britain in 36 minutes, 57 seconds, identifying 2,369,773 face-blocks that seemed to connect into 263,051 connected network neighbourhoods. However, because the analysis was done pixel by pixel, those face-blocks which overlapped two or more pixels would be considered part of different networks for every different pixel they were in. Therefore a secondary processing step was required, to identify and connect networks separated only by the boundaries of pixels. The most efficient method I could find to do this was to treat the `road_id` of the face-block street segment and its pixel-dependent `community_id` as nodes in a graph, and let `networkx` treat the dataframe of face-block communities as an edge-list, from which it can efficiently create a graph and find the connected components. This second processing step then took just 32.6 seconds, meaning the entire analysis of Britain's connected face-block neighbourhoods took less than 38 minutes.

4.3. Boundaries: Streets Connect but (Major) Roads Divide

Finally, these connected networks can then be divided by natural social boundaries, such as major roads and railways. Grannis (2009) makes more of this than I have done, distinguishing between networks which include the connecting intersections of major roads, and networks which do not intersect with any major roads. The former he terms *islands*, the latter *communities*.

Partly this is because I operationalize the idea of face-block networks somewhat differently: whereas Grannis works closely from the official designations of roads as primary, secondary or tertiary (which I suppose correspond loosely to the British designations of 'A Road,' 'B Road,' and 'Local Road'), and includes all tertiary streets as residential neighbourhood streets, I have attempted a more precise data-driven definition of what counts as a residential street.

Figure 8: *Additive Boundary Tessellations: by Motorways & A Roads; + B Roads; + Minor Roads*

Also, it is unclear to me whether the relevant boundaries should just be motorways and A roads, or B roads too, or even designated 'minor' (compared to A and B roads) roads also (see Fig. 8). Railways are also natural boundaries to pedestrian mobility, as are rivers. But the way

that we have approached the issue actually means that we do not really need to worry about them, because our connected networks of face-blocks will only cross them if there is in fact a residential street segment crossing the railway or river – and in the presence of such a bridge, they would stop being such a boundary to pedestrian travel.

Nevertheless, once we have chosen what we think the boundaries should be, we can create tessellated boundary tiles by performing a unary union on all the relevant boundaries, and then polygonizing. Then we can restrict our selection of face-blocks in a given connected network to those inside the tile, and since it is possible that the part inside the tile was only connected by street segments now outside the tile, we would need again to find the connected components of the street network graph.

5. Conclusion

Our aim in this paper was to establish a rigorous ontological basis for geodemographic analysis, by investigating and defining a meaningful neighbourhood unit. While the administrative geometric polygons sometimes referred to as neighbourhoods are generally arbitrarily modifiable areal units, the pursuit of a theoretically-grounded answer to the question of *who is my neighbour* opens up a "true causally relevant geographic context" in which such relationships necessarily develop. Specifically, the micro-geodemographic *natural area* of the "face-block," and also to the network of connected neighbouring face-blocks.

Further work must now proceed to establish how to transform demographic neighbourhood data from the arbitrary areal units with which it is associated, to the topological networks we have identified as meaningfully representing neighbourhoods.

Bibliography

Abbott, A. (2017). *Department and Discipline: Chicago Sociology at One Hundred*. University of Chicago Press.

Alihan, M. A. (1938). *Social ecology*. Columbia University Press.

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, *49*, 45–53.

Arribas-Bel, D. (2019). A containerised platform for Geographic Data Science: gds_env.

Ashby, D. I., & Longley, P. A. (2005). Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS*, *9*(1), 53–72.

Barthelemy, M. (2017). From paths to blocks: New measures for street patterns. *Environment and Planning B: Urban Analytics and City Science*, *44*(2), 256–271. https://doi.org/10.1177/0265813515599982

Batty, M. (2021). The unpredictability of the digital revolution. *Environment and Planning B: Urban Analytics and City Science*, *48*(7), 1749–1752. https://doi.org/10.1177/23998083211043601

Birss, C. I. (2019). 77m Ltd v Ordnance Survey Ltd. High Court.

Blake, M., & Openshaw, S. (1994). GB Profiles: A User Guide.

Boeing, G. (2019). Spatial information and the legibility of urban form: Big data in urban morphology. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2019.09.009

Boeing, G. (2020a). A multi-scale analysis of 27,000 urban street networks: Every US city, town, urbanized area, and Zillow neighborhood. *Environment and Planning B: Urban Analytics and City Science*, *47*(4), 590–608. https://doi.org/10.1177/2399808318784595

Boeing, G. (2020b). The right tools for the job: The case for spatial science tool-building. *Transactions in GIS*, *n/a*(n/a). https://doi.org/10.1111/tgis.12678

Boeing, G., & Arribas-Bel, D. (2021). GIS and Computational Notebooks. *arXiv Preprint arXiv:2101.00351*.

Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79.

Booth, C. (1904). *Life and Labour of the People in London*. London ; New York: Macmillan.

Boswarva, O. (2019). The 77m Ltd case: Ordnance Survey defends its address data monopoly in the High Court. *Owen Boswarva's blog*.

Burgess, E. W. (1925). The Growth of the City: An Introduction to a Research Project. In R. Park & E. W. Burgess (Eds.), *The City*. Chicago, IL, USA: University of Chicago Press.

CDRC. (2021). CDRC Maps: Maps of UK open data. *CDRC Maps*. https://maps.cdrc.ac.uk/.

Charlton, M., Openshaw, S., & Wymer, C. (1985). Some new classifications of census enumeration districts in Britain: A poor mans ACORN. *Journal of Economic and Social Measurement*, *13*(1), 69–96.

Chaskin, R. J. (1997). Perspectives on neighborhood and community: A review of the literature. *Social Service Review*, *71*(4), 521–547.

Contributors}, {GDAL. (2021). GDAL - Geospatial Data Abstraction Library. Open Source Geospatial Foundation.

Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012). Social coding in GitHub: Transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 1277–1286). Seattle, Washington, USA: Association for Computing Machinery. https://doi.org/10.1145/2145204.2145396

Dalton, C. M., & Thatcher, J. (2015). Inflated granularity: Spatial "Big Data" and geodemographics. *Big Data & Society*, *2*(2), 2053951715601144. https://doi.org/10.1177/2053951715601144

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. In *6th Symposium on Operating Systems Design and Implementation Technical Paper*.

DeReu, J. A., & Robbin, J. E. (1981). *Application of geodemographics to the Army recruiting problem*. CLARITAS CORP ARLINGTON VA.

Evans, M. (1998). From 1086 and 1984: Direct marketing into the millennium. *Marketing Intelligence & Planning*, *16*(1), 56–67.

Farr, M., Wardlaw, J., & Jones, C. (2008). Tackling health inequalities using geodemographics: A social marketing approach. *International Journal of Market Research*, *50*(4), 449–467.

Field, K. (2010). Politics and Cartography Collide: Mapping the Changing Landscape of Ordnance Survey. *The Cartographic Journal*, *47*(1), 7–11. https://doi.org/10.1179/000870410X12628661160579

Fleischmann, M., & Arribas-Bel, D. (2021). Urban Grammar research project. https://urbangrammarai.github.io/.

Fleischmann, M., Feliciotti, A., & Kerr, W. (2021). Evolution of urban patterns: Urban morphology as an open reproducible data science. *Geographical Analysis*.

Fleischmann, M., Feliciotti, A., Romice, O., & Porta, S. (2020). Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale. *Computers, Environment and Urban Systems*, *80*, 101441.

Fleischmann, M., Romice, O., & Porta, S. (2020). Measuring urban form: Overcoming terminological inconsistencies for a quantitative and comprehensive morphologic analysis of cities. *Environment and Planning B: Urban Analytics and City Science*, 2399808320910444. https://doi.org/10.1177/2399808320910444

Gale, C. G. (2014). *Creating an open geodemographic classification using the UK Census of the Population* (Doctoral). UCL (University College London).

Gale, C. G., Singleton, A. D., Bates, A. G., & Longley, P. A. (2016). Creating the 2011 area classi-

fication for output areas (2011 OAC). *Journal of Spatial Information Science*, *2016*(12), 1–27. https://doi.org/10.5311/JOSIS.2016.12.232

Galster, G. C. (2019). *Making Our Neighborhoods, Making Our Selves*. University of Chicago Press.

GDPR. (2016). General Data Protection Regulation (GDPR) Compliance Guidelines. *GDPR.eu*. https://gdpr.eu/.

Gehlke, C. E., & Biehl, K. (1934). Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association*, *29*(185A), 169–170. https://doi.org/10.1080/01621459.1934.10506247

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. "O'Reilly Media, Inc.".

Gilbert, S. (2021). *Good Data: An Optimist's Guide to Our Digital Future*. Welbeck Publishing Group.

González-Benito, Ó., & González-Benito, J. (2005). The role of geodemographic segmentation in retail location strategy. *International Journal of Market Research*, *47*(3).

Goss, J. (1995). "We Know Who You Are and We Know Where You Live": The Instrumental Rationality of Geodemographic Systems. *Economic Geography*, *71*(2), 171–198. https://doi.org/10.2307/144357

Gould, P., & White, R. (1974). *On Mental Maps*. Penguin.

Grannis, R. (2009). *From the Ground Up: Translating Geography into Community through Neighbor Networks*. Princeton University Press.

Greene, A. R., & Gilbert, S. J. (2020). *More Data, More Power? Towards a Theory of Digital Legitimacy* (SSRN Scholarly Paper No. ID 3773898). Rochester, NY: Social Science Research Network.

Greenwood, L., & Brandwood, S. (2020). Identifying properties and streets in government data - Technology in government. *Technology in Government (GOV.UK)*.

Hadden, J. K., & Borgatta, E. F. (1965). *American Cities: Their Social Characteristics*. Rand McNally.

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting*. West Sussex, England ; Hoboken, N.J: Wiley.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Vintage Books.

King, M. L. (1968). I Have a Dream. *Negro History Bulletin*, *31*(5), 16–17.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.

Klauser, F. (2013). Through Foucault to a political geography of mediation in the information age. *Geographica Helvetica*, *68*(2), 95–104. https://doi.org/10.5194/gh-68-95-2013

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., … Corlay, S. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB* (pp. 87–90).

Korner, T. W. (2013). Metric and Topological Spaces. Department of Pure Mathematics and Mathematical Statistics, Cambridge University.

Kwan, M.-P. (2012). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers*, *102*(5), 958–968. https://doi.org/10.1080/00045608.2012.687349

Lai, J., Lansley, G., Haworth, J., & Cheng, T. (2020). A name-led approach to profile urban places based on geotagged Twitter data. *Transactions in GIS*, *24*(4), 858–879. https://doi.org/10.1111/tgis.12599

Lilley, B. (2011). The ordnance survey openData initiative. *The Cartographic Journal*, *48*(3), 179–182.

Lipitz, M. L. (1969). The territorial imperative. *Public Administration Review*, *29*(4), 384–386.

Longley, P. (2005). Geographical Information Systems: A renaissance of geodemographics for public service delivery. *Progress in Human Geography*, *29*(1), 57–63. https://doi.org/10.1191/0309132505ph528pr

Louf, R., & Barthelemy, M. (2014). A typology of street patterns. *Journal of The Royal Society Interface*, *11*(101), 20140924. https://doi.org/10.1098/rsif.2014.0924

Lynch, K. (1960). *The Image of the City* (Vol. 11). MIT press.

Martin, D., Nolan, A., & Tranmer, M. (2001). The Application of Zone-Design Methodology in the 2001 UK Census. *Environment and Planning A: Economy and Space*, *33*(11), 1949–1962. https://doi.org/10.1068/a3497

Mathews, P. M., & Shapiro, I. I. (1992). Nutations of the Earth. *Annual Review of Earth and Planetary Sciences*, *20*(1), 469–500. https://doi.org/10.1146/annurev.ea.20.050192.002345

Moler, C. (1986). Matrix computation on distributed memory multiprocessors. *Hypercube Multiprocessors 1986(A 87-37501 16-61). Philadelphia, PA, Society for Industrial and Applied Mathematics, 1986,* 181–195.

Momjian, B. (2001). *PostgreSQL: Introduction and concepts* (Vol. 192). Addison-Wesley New York.

Morris, D. (1967). *THE NAKED APE*. Bantam.

O'Sullivan, D., Bergmann, L., & Thatcher, J. E. (2018). Spatiality, maps, and mathematics in critical human geography: Toward a repetition with difference. *The Professional Geographer*, *70*(1), 129–139.

Office}, {Cabinet. (2020). Geospatial Commission Charter.

Openshaw, S. (1983). The Modifiable Areal Unit Problem. Geo Books.

Openshaw, S. (1985). Rural Area Classification Using Census Data. *Geographia Polonica*, (51), 285–299.

Openshaw, S. (1997). The truth about Ground Truth. *Transactions in GIS*, *2*(1), 7–24. https://doi.org/10.1111/j.1467-9671.1997.tb00002.x

Openshaw, S., Cullingford, D., & Gillard, A. (1980). A Critique of the National Classifications of OPCS/PRAG. *The Town Planning Review*, *51*(4), 421–439.

Openshaw, S., & Gillard, A. A. (1978). On the stability of a spatial classification of census enumeration district. In P. W. J. Batey (Ed.), *Theory and Methods in Urban and Regional Analysis* (pp. 101–119). London: Pion.

Openshaw, S., & Openshaw, C. (1977). *Artificial Intelligence in Geography*. Wiley.

Park, R. (1925). The City: Suggestions for the Investigation of Human Behaviour in the Urban Environment. In R. Park & E. W. Burgess (Eds.), *The City*. Chicago, IL, USA: University of Chicago Press.

Petrović, A., Manley, D., & van Ham, M. (2020). Freedom from the tyranny of neighbourhood: Rethinking sociospatial context effects. *Progress in Human Geography*, *44*(6), 1103–1123. https://doi.org/10.1177/0309132519868767

Pickles, J. (Ed.). (1995). *Ground Truth: The Social Implications of Geographic Information Systems* (1 edition). New York: Guilford Press.

Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017). Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 1–2). https://doi.org/10.1109/JCDL.2017.7991618

Reibel, M. (2011). Classification Approaches in Neighborhood Research: Introduction and Review. *Urban Geography*, *32*(3), 305–316. https://doi.org/10.2747/0272-3638.32.3.305

Ricercar. (2021). Jonathan Robbin: Curriculum Vitae.

Robbin, J. E. (1980). Geodemographics: The New Magic. *Campaigns and Elections*, *1*(Spring), 25–46.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357. https://doi.org/10.2307/2087176

Russell, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach, eBook, Global Edition*. Pearson Higher Ed.

Schelling, T. C. (1969). Models of Segregation. *The American Economic Review*, *59*(2), 488–493.

Schirmer, P. M., & Axhausen, K. W. (2016). A multiscale classification of urban morphology. *Journal of Transport and Land Use*, *9*(1), 101–130.

Shevky, E., & Bell, W. (1955). *Social Area Analysis: Theory, illustrative application, and computational procedures*.

Simmel, G. (1908). Das Geheimnis und die geheime Gesellschaft. *Soziologie. Untersuchungen über Die Formen Der Vergesellschaftung*, 256–304.

Singleton, A., & Cheshire, J. (2021). How England's complicated political geography is confusing coronavirus rules. *The Conversation*. http://theconversation.com/how-englands-complicated-political-geography-is-confusing-coronavirus-rules-152036.

Singleton, A. D., & Longley, P. A. (2009). Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science*, *88*(3), 643–666.

Singleton, A. D., Spielman, S., & Brunsdon, C. (2016). Establishing a framework for Open Geographic Information science. *International Journal of Geographical Information Science*, *30*(8), 1507–1521. https://doi.org/10.1080/13658816.2015.1137579

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, *21*(1), 3–8.

Survey, O. (2019). Open Map - Local (Product Guide). Ordnance Survey.

Suttles, G. D. (1972). *The Social Construction of Communities* (Vol. 728). University of Chicago Press Chicago.

Swanlund, D., & Schuurman, N. (2019). Resisting geosurveillance: A survey of tactics and strategies for spatial privacy. *Progress in Human Geography*, *43*(4), 596–610. https://doi.org/10.1177/0309132518772661

Szczepański, M. (2020). Is data the new oil? European Parliamentary Research Service.

Thatcher, J., O'Sullivan, D., & Mahmoudi, D. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space*, *34*(6), 990–1006. https://doi.org/10.1177/0263775816633195

Tryon, R. C. (1939). *Cluster Analysis*. Ann Arbor: Edwards Brothers.

Tryon, R. C. (1968). Comparative Cluster Analysis Of Social Areas. *Multivariate Behavioral Research*, *3*(2), 213–232. https://doi.org/10.1207/s15327906mbr0302_6

Tryon, R. C., & Bailey, D. E. (1966). The BC Try Computer System of Cluster And Factor Analysis. *Multivariate Behavioral Research*, *1*(1), 95–111. https://doi.org/10.1207/s15327906mbr0101_6

Tukey, J. W. (1977). *Exploratory data analysis*.

Vickers, D., & Rees, P. (2007). Creating the UK National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 379–403. https://doi.org/10.1111/j.1467-985X.2007.00466.x

Vincenty, T. (1975a). Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review*, *XXIII*(176). https://doi.org/10.21236/AD0657591

Vincenty, T. (1975b). Geodetic Inverse Solution Between Antipodal Points. DMAAC Geodetic Survey Squadron.

Webber, R. (1977). *An Introduction to the National Classification of Wards and Parishes*. London: Planning Research Applications Group.

Webber, R. (2006). How parties used segmentation in the 2005 British general election campaign. *Journal of Direct, Data and Digital Marketing Practice*, *7*(3), 239–252. https://doi.org/10.1057/palgrave.dddmp.4340529

Webber, R., & Burrows, R. (2018). *The Predictive Postcode: The geodemographic classification of British society*. Los Angeles: SAGE.

Webber, R. J. (1975). Liverpool Social Area Study 1971

Data: Final Report. Planning Research Applications Group.
Webber, R. J. (1978). Making the Most of the Census for Strategic Analysis. *The Town Planning Review*, *49*(3), 274–284.
Webber, R. J. (1980). A Response to the Critique of the OPCS/PRAG National Classifications. *The Town Planning Review*, *51*(4), 440–450.
Wells, P. K. (2021). Unlike other countries the UK is not addressing a vital part of its future. *PeterK-Wells.com*.
Wells, W. D. (1975). Psychographics: A Critical Review. *Journal of Marketing Research*, *12*(2), 196–213. https://doi.org/10.2307/3150443
Wirth, L. (1938). Urbanism as a Way of Life. *American Journal of Sociology*, *44*(1), 1–24.
Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, *30*(1), 75–89. https://doi.org/10.1057/jit.2015.5
Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (Main edition). Profile Books.