Locating Spatial Data in the Social Sciences

Dr. Jonathan Reades

16 June 2021

Bio: Jon is an Associate Professor at UCL's Centre for Advanced Spatial Analysis in London. His work focusses on the application of Geographic Data Science methods to problems in housing and neighbourhoods, knowledge work and innovation, and 'smart cities' and 'big data'. An undergraduate degree in Comparative Literature has imparted an appreciation of the challenges faced by those learning programming and data analytics from scratch, and this informs not only his teaching, but also his approach to 'open access' content.

Abstract: This chapter is intended as an accessible introduction to the kinds of practical issues that social scientists can expect to encounter as they begin to work with spatial data. Taking as a starting point a single tweet, we explore the different ways that location can be embedded in data, with different 'clues' often pointing in wildly different directions! I stress the importance of slow, careful and, above all, critical engagement as the key to unlocking the power of these new forms of data. The larger the data set the more important this engagement becomes: we should never confuse 'volume' with 'truth', but provided that we frame our questions and our findings with care, that also shouldn't stop us from making the most of the exciting opportunities for social science research made possible by our always-on, digitally-enabled lives.

Keywords: spatial data, geo-data, gazetteers, address, neighborhoods, shapefiles, open street map, maup.

Introduction

The past twenty years have seen a transformation of the social sciences: data has gone from being hard to collect at scale to being seemingly 'open and everywhere' (Arribas-Bel, 2014). The past decade has been a good time to be a computationally capable researcher—as the movement of computer scientists and physicists into the social sciences has demonstrated (O'Sullivan & Manson, 2015)—but a challenging time to be a 'classically trained' social scientist trying to make sense of this brave new world of 'big data' and machine learning.

Moreover, a lot of this data has also been, as Arribas-Bel *also* noted, 'accidental' in the sense that it was never really designed to support robust (spatial) analysis. Many of these new forms of data are behavioural—generated as a byproduct of human activity such as phone calls or travel-card use (*e.g.* Reades, Zhong, Manley, Milton, & Batty, 2016)—meaning that they are rarely, if ever, straightforward to collect, organise, and interpret; but they also promise to help us bridge what has long been a critical gap in the social sciences: what do people *actually* do when social scientists aren't watching them?

However, the starting point for such work is often just working out how to deal with the data in the first place: how do we extract and store it, project and map it, and interpret or analyse it? These issues that are rarely, if ever, covered in a bog-standard 'statistics for social scientists' class where the data sets are often pre-cleaned, pre-packaged, and pre-interpreted without any reference to geography or to the kinds of problems that you might encounter as a student undertaking Independent research for the first time… or the fifth.

So this is actually a chapter about questions, not answers. I hope to show you that there is no 'right' way to tackle geospatial data analytics but, rather, a series of choices that need to be articulated, made, and documented based on the interactions between the research question and the data. Through the lens of a tweet, we'll consider the various ways that 'location' can be extracted from the sorts of data that social scientists use in order to tackle the kinds of questions that social scientists ask.

We will begin by looking at the basic classes of spatial data—points, lines, and polygons—and consider some of the most common challenges that a social scientist might expect to encounter in handling coordinate data. We then turn to locational data embedded in free-form text since, although it's often overlooked, it is likely to be encountered in archival work as well as when trying to perform 'straightforward' (i.e. not straightforward at all!) address matching. These point us towards more subtle questions around relationships—across time as well as space—between observations.

A section on temporal analysis and the deceptively simple problem of how to define a 'neighbourhood' rounds out the fundamental conceptual challenges facing the novice analyst and positions us to think about other, more practical challenges. So the the fourth section considers some of the nitty-gritty of how we access and disseminate spatial data so that we can both build on the work of others *and* share our own work in ways that support others. Finally, there is a gentle introduction to the kinds of analytical challenges that spatial data present when we fail to consider the *how* of data generation and the *why* of statistical analysis.

By the end of this chapter, I hope that you will have had a gentle (if rapid) introduction to key concepts and terminology in spatially-enabled social science, have developed an understanding of the range of contexts in which spatial data can be located (pardon the pun), and developed a basic appreciation of the analytical challenges posed by such data.

Where's Wally?

So with these ideas in mind, let's start with a seemingly straightforward question: how many 'bits' of spatial data are contained in the tweet below?

> "Love this photo of Central Park #I❤NYC" — @jreades; *Location*: Walthamstow, GB; *Geo-tag*: 43.653°N, 79.383°W; Tweeted at 01:30 (GMT-5).

Why don't you take a minute to write some down? I make it that there are *at least* four, and we'll tackle each in turn to see how they shed light on the issues I've outlined above. A more philosophical, and less data-focussed, version of this approach can be found in Crampton et al. (2013).

Working with Coordinates

The most obvious piece of spatial information in the tweet is the coordinate pair from the geo-tag: `43.653°N, 79.383°W`. Being supplied with *explicitly* spatial data often seems ideal because it gives the impression that no further thinking is required: *this* is where the event happened, end of story. In fact, the decimal latitude and longitude are for a point in Toronto, Canada—near City Hall and Nathan Phillips Square if you must know—which featured nowhere in either the text of the tweet or it's other metadata (*i.e.* the user-specified location).

We've not even begun to think about the content of the tweet and *already* we need to decide if a tweet about New York is relevant if it's sent from Toronto! So while it's true that Latitude, Longitude, and Elevation can theoretically be used to uniquely locate any point or event on the planet, as soon as humans get involved things get a lot more complicated.

We have a nasty habit of assuming that spatial data from computers is somehow 'true': when we see latitude and longitude to four, six, or eight decimal places it seems… accurate. We are then minded to accept these very precise numbers as 'truth' when, in reality, eight decimal places would imply that our GPS was accurate to within 1mm! It isn't, not by a long shot, but that sense of 'the (big) data is always right' is presumably why people still follow their GPS into the river despite the many warning signs *en route*?

So this is the simplest possible representation of this data—a point in space—and we are *still* forced to make choices about 'meaning'! One plausible choice that a researcher could make is to focus solely on tweets that are made *within* the five boroughs of New York City; we could then use this to determine whether there are hotspots of tweeting activity and try to connect these to aspects of New York life… But note that that eminently sensible decisions leads to *this* tweet being dropped from the analysis even though it's clearly about Central Park in New York City.

Get to the Point: Types of Spatial Data

The point is only *one* of the three basic building blocks of spatial data; the second is the line. A line is composed of two (or more) points in a sequence: if you were drawing a line in a notebook then there would *always* a 'first' point and a 'second' point even if your drawing has no meaningful direction. It's the same in a spatial data set.

So another option would be to treat this tweet as one point in a *sequence*: we could link up all of the points where @jreades tweeted as a set of lines (each consisting of a pair of points). Doing so, we could create a travel history for @jreades and get a sense of how he, or she, moves about regardless of whether or not they live in New York. And a polygon—which is a sequence of lines where the end point of the *last* line is the same as the starting point of the *first* one—could be drawn around *all* the points where @jreades tweeted (the technical term would be "Convex Hull") to give a sense of their 'territory' or 'range'.

So far, so straightforward, even if these might seem like a fairly poor ways of capturing the many rich ways in which humans experience and represent space. But each of these classes—points, lines, or polygons—may *also* be single- or multi-part: a single-part

geometry is one where each feature has its own attributes, and a multi-part geometry is one where several features that we see as separate 'things' on the map all share the same attribute. The polygon delimiting New York City's boundaries is a multi-part geometry because Manhattan and Staten Island are not continguous with the other three boroughs. It's the same for London because North and South London are separated by the Thames.

Rather less obviously, you might *think* that a road is a good candidate for a single-part feature because we know from driving along roads that they are continuous 'things'… However that Britain's A11 highway is multi-part feature because some stretches have been upgraded to a motorway (and renamed the M11) while other stretches have not! Furthermore, a highway that is best represented as a line at one scale (zoomed out) might be better represented as a set of lines, or even polygons, at another (zoomed in) because it is actually two carriageways with multiple lanes separated by a meridian.

Social scientists will come face-to-face with representational problems when working with geospatial data because the real world is much messier than the computer's tidy world of points, lines, and polygons: London's Paddington Station is actually made up of *two* Tube stations linked by a rail station. So if we looked at smart card data from Transport for London and just totalled up tap-ins and tap-outs at 'Paddington Station' then we might seriously over-estimate the number of people using this interchange: some passengers will be double-counted when all they are doing is transferring between lines!

How your data is generated and how you choose to represent it *matters*, so it's important to think about what you're trying to achieve when working with geospatial data: do you need to know *where* racially-motivated crimes happen within a city, or only *that* they happen within a city? The difference between those two questions might structure *everything* about your work: what methods, what types of statistical analysis you can perform, and what kinds of findings you can present. Thinking more deeply about the role that space plays in your work is critical to undertaking good social science research, and there are complex feedback effects between the types of data collected and the types of knowledge generated.

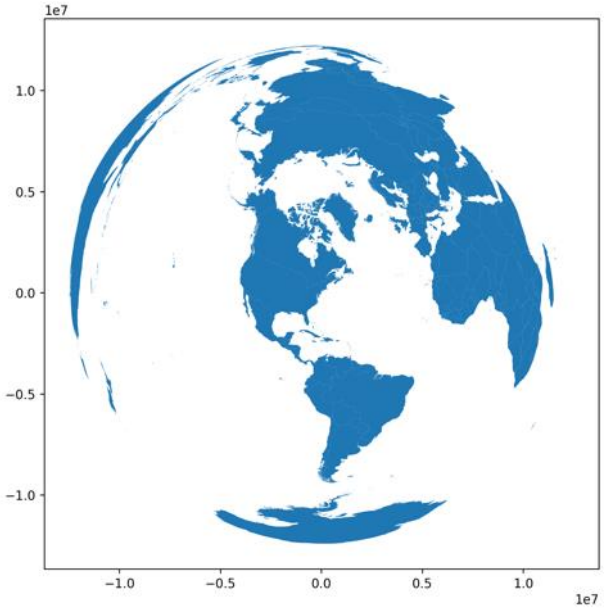
You're Projecting: Going from Round to Flat

The lat-long coordinates in @jreades' tweet were most likely generated from the phone's GPS, in which case they were 'recorded' using the top-notch WGS84 standard: the World Geodetic System reference (1984) which uses data about the Earth's gravitational and magnetic fields to locate you anywhere on the planet with astonishing accuracy. But rather confusingly for the novice researcher, you don't *make* maps in WGS84, and that's because the world is round and your map is flat.

Conceptually, it boils down to this: take a ball and then try to wrap a piece of paper around the ball without leaving any creases or folds. You might be able to wrap the paper around *part* of the ball without a crease but you certainly won't be able to wrap it the *whole* way around. There are many different ways to wrap the paper around the ball: we can start in different places so that the creases appear in different areas; we can try to minimise the total number of folds but have some big ones at the edges; or we could have lots of very,

very small folds so that there are folds everywhere, but no big distortions of the paper *anywhere*.

The more of the Earth that we want to show on a map, the more approximate the mapping (the more 'creases' we need). Below are two world maps: both are accurate, but it should also be obvious that they are both also in some sense, wrong. So a world map is everywhere inaccurate to *some* extent, and because of this many countries developed their own *projections*: these are more accurate *within* a country (fewer creases in one area) at the price of being much less accurate everywhere else. For instance, in Great Britain, we typically use the British National Grid (EPSG:27700), but since America is a rather larger country there are multiple projections because New York City (New York Long Island; EPSG:2263) is a long way from San Francisco (San Francisco CS13; EPSG:7132) or Honolulu (Hawaii zone 3; EPSG:6633).



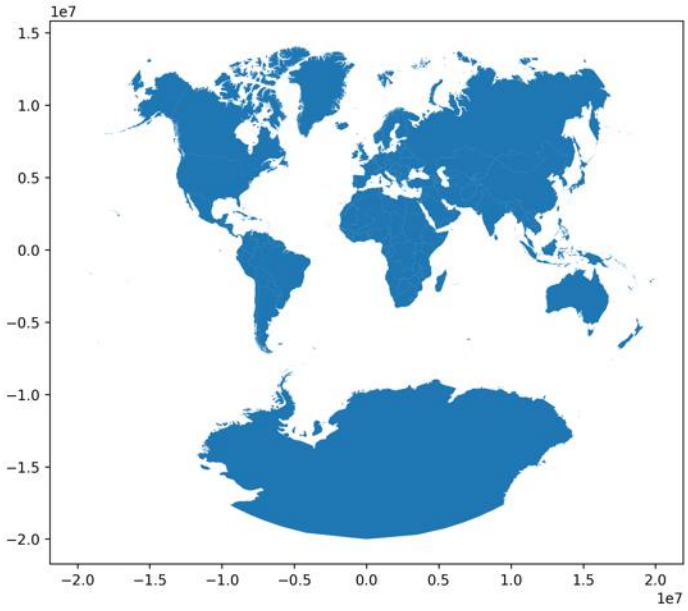


Figure 1A. Lambert Azimuthal Equal Area Projection (Data CC BY-SA 3.0, Bjorn Sandvik 2009)

Figure 1B. Ven der Grinten Projection (Data CC BY-SA 3.0, Bjorn Sandvik 2009)

You'll notice that, after each example, I listed an EPSG number: EPSG stands for the European Petroleum Standards Group and every widely-recognised projection will have its own, unique EPSG number so that two mapmakers can be confident that they are handling their data in the same way. The origins of spatial data are strongly associated with natural resources management and extraction: perhaps you can imagine why these firms would be interested in accurate geospatial data?

Below is an example of the world 'mapped' in QGIS (2021) using the British National Grid (BNG); notice the distortion of North and South America, and the way that parts of Asia are wrapping around the edges of the map to create large blocks of blue. The BNG projection becomes less accurate as you move away from Great Britain: the mathematical transformations that ensure the accuracy of data in the vicinity of the British Isles start to

break down. In addition, it's worth noting that BNG uses metres even though speed limits and distances are usually specified in miles.

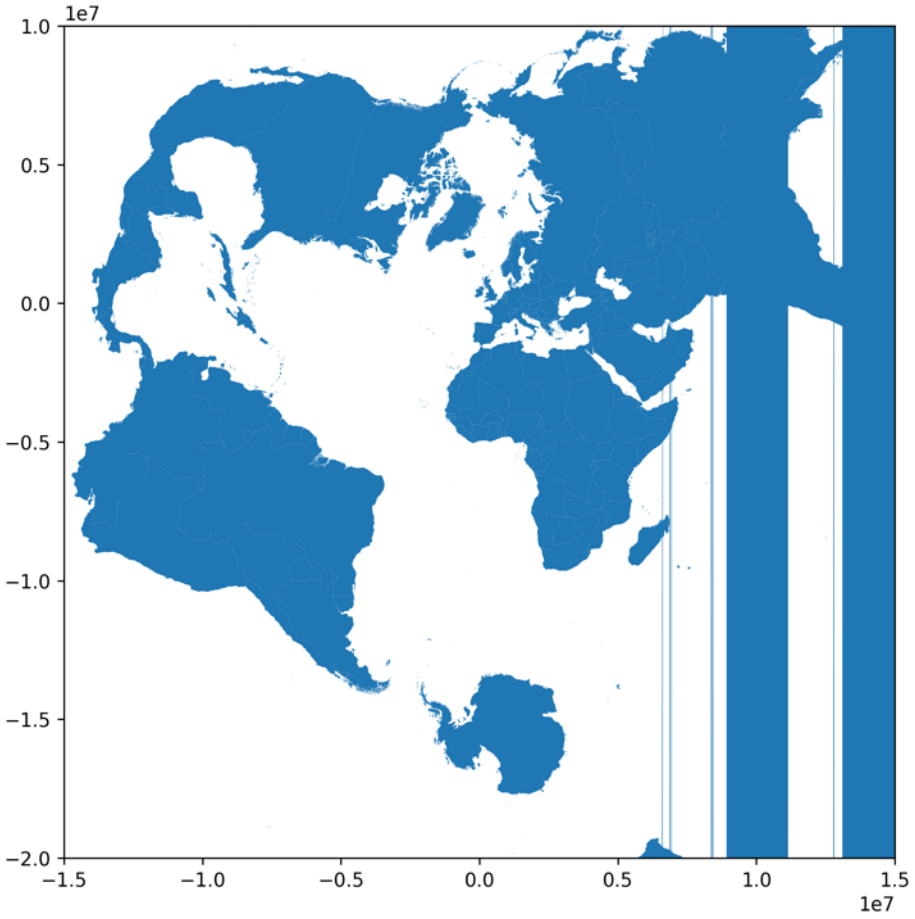


Figure 2. The World in British National Grid (Data CC BY-SA 3.0, Bjorn Sandvik 2009)

So if you tell your GIS application that the spatial data you've just loaded is in one projection or coordinate system when it's *actually* recorded in another, then your data can end up in the wrong part of the world, or even not on the map at all! Worse, projections aren't just about the part of the world you want to map, it's also about the units used to record the data: some countries record locations using metres and some use miles… and some, like Great Britain, use both! So you can also easily encounter issues of scale as well as location: you think something's in kilometers, but it's actually in degrees or miles!

Loading spatial data and seeing nothing at all on the map is the fastest way to end up thinking that geo-data and spatial analysis isn't worth the hassle, but if you can bear to deal with—and debug—problems with projections, then it's often smooth(er) sailing ahead!

Working with Text(s)

In addition to the coordinate data embedded in the geo-tag, we also had *two* other types of locational data in @jreades' tweet: 1) the *Location* specified by the user in their Twitter profile; and 2) the reference to Central Park in the body of the tweet. I distinguish between

these two pieces of text because in the first case we *know* from the field name ('Location') that the text should contain some kind of spatial identifier, whereas in the second case we have no way to be sure (without reading the tweet) that any geographical information will be found. For a single tweet this isn't an issue, but if we're trying to find locations in a book or in a collection of tweets amassed over a period of months, then we will need to enlist the help of a computer.

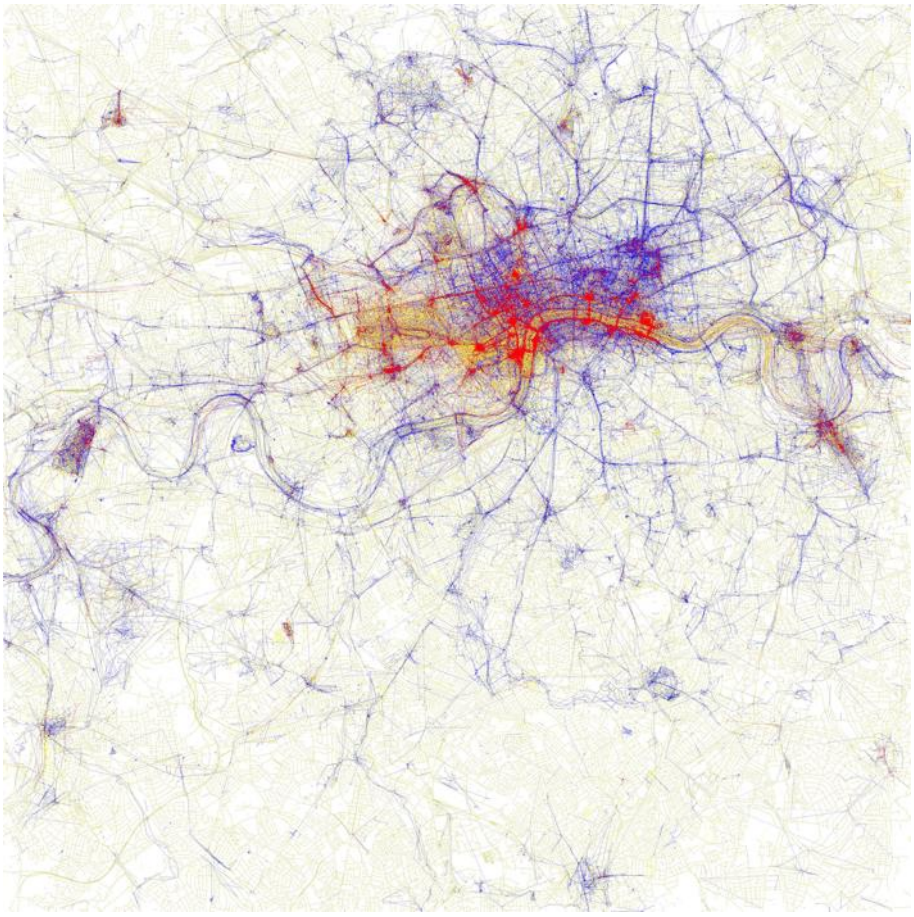


Figure 3. London, Locals and Tourists (CC-BY-SA, Eric Fisher 2010)

In the map of London shown here, Eric Fisher has contrasted the geo-tagged coordinates of photos posted to Flickr with the home location specified by the users who uploaded them: blue for locals, red for tourists, and yellow for undetermined. But note that, unlike coordinate data which is unambiguous—even if easy to misinterpret—with textual data on a 'home location' the location could be missing (yellow), incomplete (yellow), mis-spelt (yellow), in a 'slang' form (yellow), or even indeterminate because there are multiple Delhis (there's one in Canada) or Berlins (there *used* to be one in Canada) !

Look it Up! The Value of Gazetteers

Knowing which bit of text contains locational information simplifies things enormously because we can be a lot 'stupider' in our approach: although the data may vary in its intelligibility and accuracy, for the type of work that Eric did above the fact remains that

there are only 200-odd countries in the world (depending on what counts as a country!). For a computer, checking short snippets of text against a small-ish dictionary of valid country names and abbreviations is easy: it can probably do tens of thousands of lookups every second when there are only a few hundred entries to check!

Manually assembling a short list of valid country names is fairly easy, but what about tens of thousands of place names? For social scientists working with text, the term 'gazetteer' will become a familiar one. Gazetteers are often produced by crowd-sourcing data on place names, and they typically provide one or more lookup 'names' (sometimes in multiple languages or multiple spellings) against which a piece of text can be matched in order to place it not only on a map, but also within some kind of formal ontology: a town, a national forest, and other commonly-used classes. This would likely be the best way for the *computer* to work out that Central Park is a location in New York City, and that Walthamstow is an area within the London Borough of Waltham Forest.

With unstructured data—such as a book or document—gazetteers can be an essential part of the mapping process, and they are particularly useful when dealing with historical data incorporating places, or spellings, that no longer exist (Walthamstow began life as 'Wilcumestowe', the Place of Welcome). Gazetteers allow us to put historical places on contemporary maps, but they also allow us to put these places in a spatial and temporal context—via linked entities or entries—that incorporates useful details such as an historical parish or other administrative unit. For instance, the entry for Skara Brae in Scotland looks something like this:

Name	**Skara Brae**
Type	Archaeological Site
Built	3100 BC
Text of Entry Updated	05-FEB-2016
Latitude	59.0492ºN
Longitude	3.3456ºW
National Grid Reference	HY 229 188
Source	Clarke, David (2000) Skara Brae. Historic Scotland, Edinburgh
Linked Entries	St. Peter's Church, Orkney, Sandwick, Douby
Features	Orkney Mainland; Skaill House; Skaill, Bay of

Table 1. The Gazetteer for Scotland (Gittings, n.d.)

Each of these items can, in turn, be linked to other entries in the gazetteer for additional information, allowing us to convert plain-old text into something much richer and more structured, though see (Gittings, 2009) for both perspective on the value of (geo)text as well as a fuller discussion of the challenges entailed in regularising something as 'simple' as a place name. Aside from the impossibility of recording information about *everything*, *everywhere*, gazetteers have tended—historically, at least—to represent most features as points since that made them easy to download as flat text files. This limitation is no longer

relevant, but many gazetteers are still presented in this way and we are (again) back into questions of the *approrpriate* representation of spatial features in data.

Locational data can be found embedded in *all* forms of structured and unstructured text: web pages, letters in archives, the institutions associated with PhD theses… The location might be an actual address, a town or village, an institution or crossroads, but if we can distinguish these spatial 'terms' from the rest of a document then we can begin to treat them as geographical data and start to ask spatial questions! One area where I expect to see significant change in the near future is in the growth of geographically-aware Natural Language Processing (NLP) libraries that are able to extract spatial information directly from source texts without relying quite so heavily on what are basically enormous dictionaries.

Answers on the Back of a Postcard: Using Address Data

Gazetteers might seem clunky if you've got access to address data! However, the latter presents *other* challenges to the social scientist, particularly where historical data is concerned. Some time ago I was approached by a political economist wanting to map Soviet-era factory data that researchers had painstakingly collected and systematised with a view to better-understanding how the communist production system worked. The challenges here were manifold: since the Soviet Union no longer exists, many of the places no longer do either—city names have been changed, streets no longer refer to heroes of the Marxist-Leninist Revolution, and the factories have long-since ceased operation—and there is no contemporary source of data against which to validate these locations.

Worse, the addresses had been inconsistently recorded: abbreviations, incomplete postcodes, multiple addresses for a single factory, multiple factories (clearly in different cities) stored with a single address field… and that is before you consider the inevitable data entry errors resulting from the fact that it was input by human beings. *None* of this is to criticise the enormous and admirable amount of work that went into collecting such data; rather, I wish to highlight the challenges that can be expected: to imagine that 'address matching' is a straightforward process that can be 'tacked on' later is to set yourself up for frustration, and possibly failure. Should we abandon this effort? Absolutely not. Should we be rather more conservative in our estimates about the amount of time and energy that will be consumed by 'getting the data into shape'? Absolutely.

A more subtle consideration is that this kind of data is often only useful if you have access to the appropriate reference information. And access to such reference data can be a big *if*: in the U.K., for instance, when the Government privatised the Royal Mail it also, rather carelessly, privatised the Postal Address File that is the canonical source of all address-related information. Suddenly, thanks to weak thinking about data governance, the *real* boundaries of postcodes as well as the *exact* location of Harry Potter's 4 Privet Drive, Surrey (actually 12 Picket Post Close, Berkshire), were a source of additional income *and* under the control of a for-profit entity.

Rather unsurprisingly, the company is not particularly interested in making such commercially valuable data readily available to the world. As an academic, I have a license to access the postcode *boundary* files, but I cannot publish any data that I *derive* using those

boundaries without falling foul of the licensing terms. The effect of this is to 'pollute' the analytical pipeline in a way that makes the re-use of address data for other purposes profoundly problematic. Bizarrely, it is actually safer for me to work with the less accurate postcode *centroids* (the inferred middle of each postcode polygon) since they are not covered by the same onerous licensing terms.

The decision on whether to use privileged data as part of an analysis is nearly always one to be taken on a case-by-case basis: if others cannot replicate your findings or make use of your code/analysis then this represents an impoverishment of the wider research landscape (including private, public, and third sector work). You may even—as I've experienced—end up on the wrong side of your *own* research: some time ago I desperately wanted to revisit some earlier work that I had done with the more sophisticated techniques that I now understand because I think that there is a lot more that I could do. But I can't, because as part of my condition of access, the data had to be destroyed once the initial work was done and the company that supplied it doesn't retain data for years.

This kind of constraint may be worth accepting when, say, the accuracy or timeliness of results is paramount. However, as a result of my own experiences I have, personally, increasingly prioritised the openness of my work and my data on the basis that this is the best way for others to learn, to critique, and to enhance. My point is not that there is no justifiable reasons for making other choices in this regard—some of my earliest work with telecoms data simply could not have been done had openness been required (Reades, Calabrese, Sevtsuk, & Ratti, 2007; Reades & Smith, 2014)—but that consideration should be given to how the data might be used now… *and* in the future.

Address-type data can also present challenges to confidentiality: Prof. Latanya Sweeney managed not only to identify the Governor of Massachusetts in an 'anonymous' health data set using nothing more than his date of birth and zip code, but has gone on to show that publicly accessible profiles in the Personal Genome Project can be deanonymised more than 80% of the time (Information, n.d.; Sweeney, Abu, & Winn, 2013)! Maintaining the privacy of sensitive records requires enormous care, and failure to understand how easily purportedly anonymous individuals can be reidentified is a major reason why social scientists can find their discussions with institutions or corporations around data access going nowhere fast. Researchers should seek to obtain the *least* resolution that still supports their analytical objectives: in spite of what I say below about the MAUP, linking individual data only to standard Census-type geographies *and* asking for attributes such as age to be grouped into bands is the *only* way to moderate this risk.

Linked In: Joining Spatial and Non-Spatial Data

The process of looking up information in other databases or data sets is a kind of data linkage: we *join* two data sets together by matching information from one data set to information in another in order to gain access to its (geospatial) features. When making a map, many of the data sources (including the Census) upon which social scientists depend come in a 'tabular' form such as an Excel or CSV (Comma-Separated Value) file that requires us to join it to another data set containing spatial information. So how do you join your spatial and non-spatial data files, and what do you need to look out for?

National statistics and mapping agencies are used to this problem so they typically provide data in a standardised format that is *designed* to be easy to link up. Below is a tiny extract of data from the 2011 U.K. Census provided by the Office for National Statistics:

2011 Super Output Area - Lower Layer	Mnemonic	All usual residents	...
Camden 001A	E01000907	1430	...
Camden 001B	E01000908	1581	...
...	...	...	...
Camden 028D	E01000919	2014	...

Table 2. Census Data Table (Source: Office for National Statistics licensed under the Open Government License v.1.0)

The two fields of interest here are the '2011 Super Output Area - Lower Layer' and 'Mnemonic': these both refer to the same spatial *feature*—a Lower Layer Super Output Area (LSOA, for short) which is similar to a Census tract in the U.S.—and then attach a number of *attributes* about that feature (*e.g.* the number of usual residents, the number of women, the number of people aged 16–24…). These two fields are unique by design because they are the *key* to joining your data set to a geodata file.

Why have two fields? The Mnemonic is a fixed-width code (*i.e.* it's always 9 characters long) so it's efficient to store and it is guaranteed to be unique. The human readable LSOA name is easier to understand, but the field must be as long as the longest Local Authority name in the UK *plus* the 4-digit identifier that ensures uniqueness. This added flexibility not only takes up more storage space on a computer (though it's not a lot for most modern computers), but it's also slower to join and harder to tell if you've made a small mistake that might have led to mis-matches. If in doubt, it's best to go with the encoded form.

In order to make a map we need to join this data file to a spatial data file containing points, lines, or polygons. Below, in rather abbreviated form, is what the matching rows from *one* type of spatial data file might look like. Notice how the `coordinates` in the example below are shown as pairs (*e.g.* `[ 528559.2, 186904.2 ]`)? These are the x and y of each point (also known as a *vertex*) in the polygon. The last pair of coordinates in each polygon are the same as the first pair because the polygon must be *closed*.

Selected JSON Features for LSOA Geometry
{ "type": "feature", "properties": { "fid":1, "lsoacd":"E01000907", …}, "geometry": { "type":"MultiPolygon", "coordinates": [[[[528920.9, 186917.1], [528935.1, 186831.6], [528940.3, 186801.8], …, [528920.9, 186917.1]]]] } },
{ "type": "Feature", "properties": { "fid": 891, "lsoacd": "E01000908", …}, "geometry": { "type": "MultiPolygon", "coordinates": [[[[528559.0, 186904.2], [528561.9, 186861.9], …, [528559.0, 186904.2]]]] } },
{ "type": "Feature", "properties": { "fid": 902, "lsoacd": "E01000919", "LSOA11CD": "E01000919", "LSOA11NM": "Camden 028D", …}, "geometry": { "type": "MultiPolygon", "coordinates": [[[[530052.3, 181554.1], [530108.0, 181450.0], [530136.0, 181456.0], …, [530052.3, 181554.1]]]] } },

Table 3. List of Features (Contains Ordnance Survey data © Crown copyright and database right 2021)

You *join* records in each data set together by telling the computer how to make matches between the two files; in this case it's the `Mnemonic` column in the data file and the `lsoacd` field in the spatial data file. Your GIS application or coding libraries will (hopefully) understand how to extract possible matching names from the format above and list the non-geometry fields as options for linking the two files.

On occasion, when working with a GIS application such as ArcPro or QGIS, or with code in Python and R, you may be asked whether a join is `1:1` or `1:n`. While daunting, this question can be better understood as: could 'things' in one data file have multiple matches to 'things' in the *other* data file? As a heuristic: statistical and other formal data releases by government agencies tend to involve `1:1` links because the data have been cleaned and organised for you, while 'messy' data collected via scraping or ad-hoc Freedom-of-Information (FoI) Requests—as well as data involving mis-matched scales—tend to be `1:n` because the relationship is not straightforward.

So if you have Census data and a Census spatial data file then the answer will nearly always be `1:1` because the files are *designed* that way: for each Census zone in there should be one, and only one, match between the data set and the spatial data set. In pratice, with `1:1` joins there is exactly zero or one matching rows for each row in the spatial data file. If more than one match is found then only the first match is kept and all other matches are (silently) discarded.

But if, for instance, you wanted to look at Airbnb listings in a densely-populated urban Census area then you'd expect there to be many matches between listings and a Census area so that would be an `1:n` join. So `1:n` joins are normally encountered when you are *aggregating* or *grouping* data together, such as when we want to calculate the total number of people living in a borough from the number of people living in each LSOA or Census tract. If you don't see what you were expecting (and it's not a projection issue) then the specification of the join is another likely culprit.

Spatial joins—also known as 'joins by location'—are a special case of `1:n` joins because we don't use a pair of columns in the data files to make the match, we use actual locations instead: we might take number of crimes (recorded as a point) and use a spatial join to calculate the crimes within each LSOA (recorded as a polygon). You will normally also need to tell the computer whether and how to calculate derived variables based on the matches it makes: *e.g*. sum the number of crimes together. It's also common to join polygons to polygons or lines to polygons, but beware: what should your GIS do if a line crosses two or more polygons? Even if it's 1mm of a 2km line that crosses over, as far as the GIS is concerned that is enough to give you a `1:n` join!

For this reason spatial joins are hard work for computers: it's much faster and more accurate to match on fields than on geometries. If you *have* to do a (big) spatial join then it's common practice to take the centroids—the 'centres' of a line of polygon—of one data set before performing a join: with individual points you're much more likely to end up with them falling inside only one polygon. Although there are clearly times when this is *not* the right strategy, because calculating points-in-polygons is *fast* there are many situations

where this is the best way to join data sets from different—of differently-scaled—geographies.

Maps are Manifolds: Incorporating Time

Finally, although it's not obviously a piece of geospatial information, `01:30 (GMT-5)` is also nonetheless inherently spatial as well. Most obviously, if we slice up the original data by date and time we can start to ask questions like: where and what do people tweet after midnight? where do they tweet after noon on Saturday, or on Tuesday? Leaving to one side questions about *who* is tweeting at 1:30 in the morning, the potential to segment behaviours—tweeting, travel, complaints, crimes—in time and space with increasing granularity is why geospatial analysts are in such demand: they can make sense of these spatio-temporal patterns and, with luck, their socio-economic context.

But there's also a deeper question that can be brought to light by asking: how far is it from Manhattan, New York to Hoboken, New Jersey? Assuming that you know where these two places are, then chances are that you instinctively came up with an answer that was measured in minutes, not miles or kilometres. Travel time is a perfectly legitimate way to measure distance, but it already contains several built-in assumptions: were you intending to travel by train or by car? At rush hour on Monday, or early on Saturday? And did you mean from the middle of Manhattan to the middle of Hoboken, or from water's edge to water's edge? In fact, if you are measuring from adminstrative boundary to adminstrative boundary then the distance between these two areas is *zero* because Manhattan and Hoboken touch in the middle of the Hudson River!

Many of us instinctively frame and understand a question about space in terms of the *experience* of a journey and not the *number of units of distance* traversed. An extreme illustration of this effect would be the so-called 'postcode gangs': for some young people in cities like London or New York leaving your hyper-local area can mean taking your life in your hands. Distance for a gang member is not remotely linear: a job opportunity a few blocks away, if it's on another gang's 'turf', might as well be on Mars. This issue should clue us in to the idea that geospatial analysis is as much about choosing the best *representation* of space from amongst many possible representations as it is about *measuring* it precisely afterwards.

Riffing on the statistician (Box, 1979), we could say that 'all models *of space* are wrong, but some are useful'. So if the question is 'how far is it from 51st and 2nd to that bar in Hoboken?' then 'It will take zero minutes to get there because the two counties are adjacent' isn't a very useful representation because our question isn't about administrative units, it's about travelling between points on a network (see, for example, the chapter by Ye, Peng, & Gong, 2021). But if we're interested in the impacts of Manhattan's real estate market on its neighbours, then the fact that Hoboken is 'adjacent' *is* profoundly relevant. Similarly, if the question is 'how far is it from the gang's home turf to the job in Pret-À-Manger' then 'it's *just* 750m' is also not a very useful representation because it ignores the way that distance is experienced by a former gang member.

There Goes the Neighbourhood: Thinking About Near and Far

This inevitably brings up the issue of how to define 'near' and 'far'. Usually, near is defined as something 'falling within the neighbourhood of the feature of interest' but that is hardly a robust definition. A slightly better definition would be the distance—using the most appropriate representation—over which we expect some kind of interaction to occur. The epidemiologist, public health researcher, gentrification researcher, or political scientist will all define this interaction distance in different ways: it could be a housing market or area with a distinct demographic profile (see, for example, the chapters by Delmelle, 2021, and @knapp:2020), an electoral ward or area experiencing a similar shift in voting intentions (*e.g.* the chapter by Wolf, 2021), the area served by a particular doctor's practice or hospital…

For example, if you are interested in commuting then your model might define a neighbourhood using mean (or median) commuting distance. If you are interested in public health then you might pick a radius around a point pollution source such as a factory or incinerator. Walkable neighbourhoods calculated for children, adults, and there elderly will look very different, as might those calculated for New Yorkers and Houstonians. In short, we are focussing on *effects*, not *units*, and your definition might draw on a review of the literature, a hypothesis to be tested, or a model of the process being studied.

Representing the neighbourhood in quantitative form can be a challenge because we need to calculate whether each and every location in the data set is near (within the neighbourhood) or far away (outside of the neighbourhood) from *every other location* in the data set. This is a matrix calculation and it becomes very expensive when our data set is large. For instance, if distances between locations are symmetric (meaning that we can assume it is as far from from B-to-A as it is from A-to-B) then for 1,000 data points there are 'just' 499,500 distances to calculate. You might be able to think of cases where distances are *not* symmetric, but the key idea is that the quantitative definition of a 'neighbourhood'—which is very closely tied to the concept of a cluster (see chapter by Helderop & Grubesic, 2021) should never be confused with the 'fuzzy' term we use in conversation with one another.

Working with Pre-Packs: Dealing with Data

I've deliberately avoided talking about file formats and their gnarly details since: a) they're a bit boring, and so b) they tend to be off-putting to readers. However, at *some* point we need to do this because it really does matter, and leaving it all for the discussion at the end of the chapter looks a lot like ending on a low, not a high. Pre-packaged data prepared for use with a GIS or programming environment most commonly come in one of three formats: Shapefile, GeoPackage, and GeoJSON.

I Hate Shapefiles (But They are Hard to Avoid)

Shapefiles are a bit like the `.docx` document or `.jpeg` photo formats: they've been around for *ever* so nearly everyone can read and write them, but nearly every professional hates doing so. For a start, Shapefiles aren't singular files! A Shapefile is *actually* composed from at least three separate files: the geometries in `.shp`, the shape index in `.shx`, and the

attributes in `.dbf`; however, you might also have a projection (`.prj`), a spatial index (`.sbx`), and metadata (`.xml`)! So if you want to share or back up a Shapefile then you need to collect *all* of those files into a Zip archive, and missing even one of them out can leave you with terminally corrupted data. In addition, each Shapefile can contain only *one* data type: points, lines, or polygons. So although the fact that many platforms can work with Shapefiles makes it seem a good default choice, today there are much better options.

The 'new kid on the block' is the GeoPackage: as the name suggests it *packages* up all of your data in a format that looks to a computer like a single file. This makes it easier to share, backup, and download, but there's a much bigger advantage: in addition to embedding the projection information, the GeoPackage can also include multiple layers with different types of geo-data, and some applications can also use it to store style information. So not only can you share your entire analysis (*e.g.* the polygon showing the boundary of your study area *and* the points that fell within it *and* the aerial photography that covers it), but if you spent a lot of time making your results look *good* then in some cases those who download your data get that benefit as well!

The GeoPackage is a powerful file format because it's basically a compressed database, but this can be overkill and it's also not very 'user-friendly' since it requires specialised software to read and write it. So a practical alternative to the GeoPackage is the GeoJSON file: this is effectively structured text (as you would have seen in the example above) intended to be easy for computers to parse *without* specialised software. GeoJSON can often be displayed directly in a web browser, and 'simple' web applications can still allow dynamic interaction with multiple layers, including panning and zooming, popups and custom iconography. What GeoJSON does *not* do well is scale: the other file formats support very large data sets, but GeoJSON cannot.

So once again, there is no 'right answer' only—as the English would say—horses for courses: if you need to share your data widely and it's *relatively* limited in scale then GeoJSON would be an excellent choice. If you want simplicity and elegance then GeoPackage should be your format. And if you want to ensure that every possible user is supported then there's still a role for the venerable Shapefile. These strengths and weaknesses clearly interact with those of the applications that make use of them: the limitations inherent in web browsers mean that the constraints of scale and complexity are encountered far sooner than with dedicated GIS applications, and those too will begin to fail long before you've 'maxed out' the capacities of dedicated command-line or database software such as GDAL and Postgres/PostGIS.

To Space and Beyond: New Sources of (Good) Data

The use of satellite and remotely-sensed data has not, historically, been part of the social scientist's toolkit, but they are becoming more useful as both the resolution of the imagery, and the power of the computers to which we have access, improve (see chapter by Arribas-Bel, Rowe, Chen, & Comber, 2021). Indeed, when we consider that systems such as LANDSAT now provide coverage dating back over 30 years one can, with care, begin to construct big picture histories of urbanisation and development even in areas where any

number of factors—ability to collect, corruption and conflict, or simply poor quality—might have undermined more traditional sources of data on people and places.

A second exciting source of data for places that, historically, were missing from our maps is OpenStreetMap (OSM), its humanitarian 'arm' (HOTOSM), and the allied 'MissingMaps' programme which emerged in the aftermath of the devastating 2010 earthquake in Haiti. This subsequently led to a number of initiatives to radically improve coverage of Africa and neglected parts of Asia for both research and aid relief purposes (see Myanmar map below). In general, I feel that this is a 'good' outcome, although a critically reflective social scientist would also recognise that the act of mapping nonetheless has a tendency to reproduce existing power relations: it is, for instance, much easier to find strip clubs and bars than baby-change stations or rape crisis clinics in OSM (Stephens, 2013; and see also Elwood & Leszczynski, 2018).

Figure 4. Missing Maps Contributions in Myanmar (CC-BY Missing Maps)

A related challenge with crowd-sourced and volunteered geographic information is that, in the absence of extensive validation and a strong ontology, human classification tends to vary in accuracy and consistency. So having mappers in one country assign land uses to features observed from space in another can produce wildly inaccurate assessments of what is being studied. In the case of HOTOSM/MissingMaps there is an explicit hierarchy—mappers and validators—and this is further followed up (ideally) by trained staff on the ground who are able to assign locally-relevant place names and feature attributes. However, since OSM does not really enforce a strong ontology or try to map these classes across languages it can be challenging to try to *compare* the distributions of many types of objects.

None of this is to minimise the achievement or utility of OSM: it may well be the *only* data source supporting open, replicable, and cross-border mapping, and in many cases it may have more information/detail than a standard map. However, neither should OSM be confused with the output of a national mapping agency and so it is, once again, a case of

horses for courses. But lest you think that you can avoid these issues by relying on data from a national organisation: the use of older, marginally less accurate geo-data by the U.S. Treasury and their *slight* misalignment with the most up-to-date geo-data provided by the Census allowed a CEO worth billions to claim a major tax-break intended for the poorest in America (Ernsthausen & Elliott, 2019)!

Discussion

Linking back to our original tweet one last time: for analytical purposes should we attribute this tweet to the account's address (London), the tweet's geo-tag (Toronto), or to the place referenced in the tweet itself (New York)? The answer is, of course, it all depends on what we want to know! This is the vital contribution that social scientists can make to such projects: although we have much less control than we once did with purposive surveys over how data from platforms like Twitter are collected, we *are* used to thinking critically about data and data collection, and about the ways in which it can (or cannot) be applied to a particular research question.

So focussing on Central Park as a landmark tells us something about the features that tourists and 'natives' associate with New York City. But the sequencing of geo-tagged tweets by accounts gives insight into mobility patterns and tourism, potentially at a global scale (*e.g.* Girardin, Calabrese, Dal Fiore, Ratti, & Blat, 2008). Or we could consider whether the account address is useful for designing overseas marketing campaigns ('the Brits like Central Park, the French like the Empire State Building'). None of these is a simple mapping between activity, location, and purpose, but all are potentially rich and entirely legitimate topics for social scientists—whether academic or otherwise gainfully employed—to tackle.

The Perils of Spatial Data

Working with geospatial data will introduce the social scientist to a variety of analytical and representational challenges that can only be briefly touched on here. However, perhaps the most important idea the reader can retain from this foundational chapter is the importance of thinking about how spatial data is generated and recorded before using it in your research. Good research requires consideration of *how* the system produces data, *who* produces it and *why*, and of *what* the analysis seeks to achieve: the social sciences are good at the who, why, and what, but often much less good at the how.

Geospatially and computationally-empowered social scientists can bridge this gap, bringing good reflective practices to an area where, all too often, data is treated simplistically as 'truth'. A few years ago I was involved in pitching an analytical platform to an international telecoms provider: we argued that, using techniques taken from planning, we could model the spatial distribution of their users with much greater precision. This would allow the company to improve their service offering by highlighting under-served areas and enhancing location-based services such as messaging about transit delays. The firm's data scientists responded: "We don't need this, we have *all the data* already and don't need a model." To them, the data spewed out of their mobile phone network *was* reality, and questions of bias, resolution, and representativity were irrelevant!

The problems of this mindset were driven home while talking to a researcher at a *different* phone company who'd decided to run some tests on how his own phone connected to their network because he'd discovered that none of the models in use had actually been validated! The analyst discovered that the pattern of connections was unlike anything they'd expected, with cells overlapping far more than anyone had realised and the behaviour of the network differing for fast- and slow-moving phones (*e.g.* while walking vs. talking on the train). These details *matter* because high-profile publications in journals like *Science* and *Nature* make liberal use of CDRs (Call Data Records) which are the logged billable 'events' triggered by a mobile phone user. But if it turns out that our understanding of how this system works is *wrong*—or, at least, incomplete—then there are potentially significant implications for the kinds of conclusions that can be drawn from such data.

The issue is not that mobile phone data is useless, it's that it is likely to be inappropriate for some purposes. Since relatively few researchers have privileged access to such data, this intersects with wider questions of verification and replication: do the conclusions hold for other networks with different consumer profiles? for other countries? even for other types of network (there is more than one!)?

The (Statistical) Perils of Spatial Data

Variations on this issue also crop up in statistics when dealing with spatial data: one of the most basic assumptions of frequentist statistics—that observations are independant of one another—is violated as soon as we start factoring in the role of space. The simplest way to think about it is: where are you *likely* to find wealthy or poor people? Near *other* wealthy or poor people respectively! In practical terms, if we think of statistics as a way of working out whether a pattern we're observing in our data is random or statistically significant, then using non-spatial statistics can lead us to see patterns where none exist because our confidence thresholds are wrong (we should *expect* to find clusters of wealth and deprivation!).

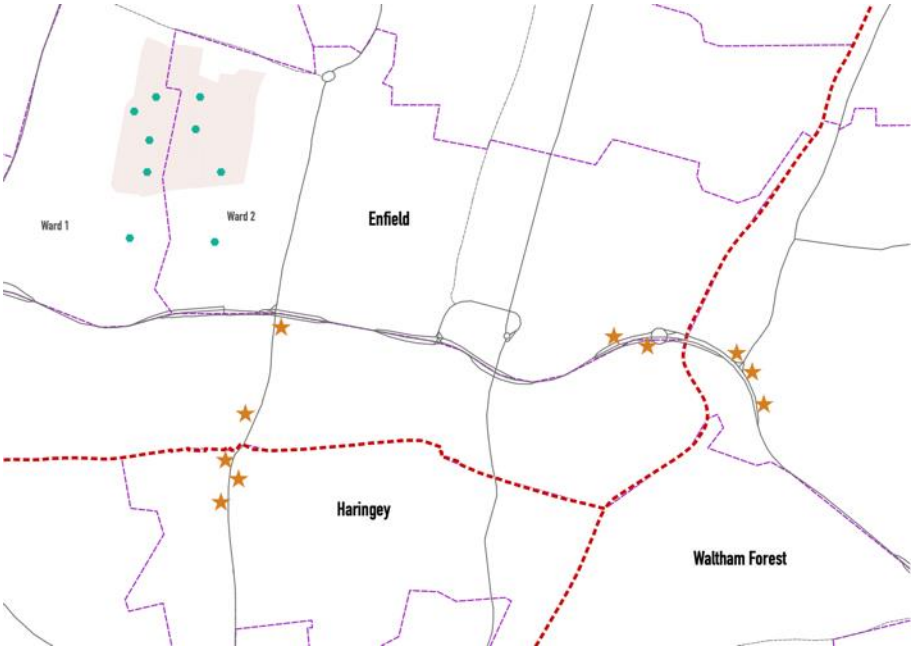


Figure 5. The Modifiable Areal Unit Problem (Contains Ordnance Survey data © Crown copyright and database right 2021)

A more subtle issue—and one about which we can, in practical terms, often do very little—is the Modifiable Areal Unit Problem (*e.g.* Fotheringham & Wong, 1991): as social scientists, we're often tempted to use administrative boundaries to help structure our analysis of, say, crime or road accidents. In the example shown here, the shaded area is obviously a hot-spot for some kind of activity (the hexagons), but the events are evenly split between Wards 1 and 2. Suddenly, events are 'dispersed' across two different, larger units of analysis! The same holds for the events recorded as stars on the map: they are clearly driven by a process connected to main roads, but a borough-level analysis would show only that Enfield had more of them than its neighbours.

A similar issue *also* happens when we try to dissaggregate data: not only is it largely (though not quite always) impossible, but it often leads to the Ecological Fallacy of inferring that, because an area is *on average* well-off, then *everyone in* that area is well-off. Similarly, although well-off *people* tend to vote for conservative/right political parties, well-off *areas* tend to vote for liberal/left parties.

The Power of Computational Thinking

Finally, it should be clear that, over the course of this chapter, we have moved into realms where hand-coding/correction is just about impossible and that to approach these problems requires code. Indeed, many of the other chapters in this *Handbook*, such as the introduction to GIScience (Buttenfield, 2021) and Analytical Environments (Bivand, 2021) are as much about the power of code to help us tackle challenging spatial problems as they are about the concepts presented; this is *because* the way that we approach these ideas is by writing code to perform them. In fact, many spatial statistics are best-presented as algorithms, not equations, for reasons that will become clear over the many subsequent chapters (for an excellent illustration of this see Xiao, 2016).

However, you can always return to the fundamental fact that geospatial data, like geospatial analyses, are produced by people and by systems created by people: there is nothing 'innate' or 'correct' or 'true' in such data—or the algorithms with which we analyse them—and you should feel free to *choose* simpler data and algorithms if you understand them and are confident that they are *appropriate*. What the rest of the book does is help you to understand how various approaches can be appropriate in particular contexts and to particular problems. Engaging critically with new concepts is integral to what we, as social scientists *already* do every day and it's how we learn to ask the questions that cut to the heart of contemporary social and policy issues.

Moreover, asking questions is critical to good (social) science: I hope that you come away from this chapter *not* daunted by the range of concepts and challenges that seem to lie in wait, but *empowered* to ask questions of others. How are these events generated and logged? Why is this the right representation of the process? What kinds of mistakes or assumptions might we be making if we assume that $n = all$? For a more in-depth tackling of these issues in a non-spatial context, the highly engaging and very accessible *Data Feminism* (D'Ignazio & Klein, 2020) would be an excellent starting point, and see also the challenges and opportunities being raised in *Part 4* of this book (Dony, 2021; Folch, 2021).

Bibliography

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

Arribas-Bel, D., Rowe, F., Chen, M., & Comber, S. (2021). The Potential of Imagery in the Social Sciences. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Bivand, R. (2021). Analytical Environments. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. Research Triangle Park, NC. Retrieved from http://www.dtic.mil/docs/citations/ADA070213

Buttenfield, B. (2021). GIScience. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating 'big data'and leveraging the potential of the geoweb. *Cartography and geographic information science*, *40*(2), 130–139. Taylor & Francis.

Delmelle, E. (2021). Neighborhood Change. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Dony, C. (2021). Computational Thinking. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Elwood, S., & Leszczynski, A. (2018). Feminist digital geographies. *Gender, Place & Culture*, *25*(5), 629–644. Routledge. Retrieved from https://doi.org/10.1080/0966369X.2018.1465396

Ernsthausen, J., & Elliott, J. (2019). One trump tax cut was meant to help the poor. A billionaire ended up winning big. *ProPublica*. Retrieved from https://www.propublica.org/article/trump-inc-podcast-one-trump-tax-cut-meant-to-help-the-poor-a-billionaire-ended-up-winning-big

Folch, D. (2021). Uncertainty. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Fotheringham, A. S., & Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, *23*(7), 1025–1044. SAGE Publications Sage UK: London, England.

Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, *7*(4), 36–43. IEEE.

Gittings, B. M. (2009). Reflections on forty years of geographical information in scotland: Standardisation, integration and representation. *Scottish Geographical Journal*, *125*(1), 78–94. Routledge. Retrieved from https://doi.org/10.1080/14702540902873881

Gittings, B. M. (n.d.). Skara brae. Retrieved 2020, from https://www.scottish-places.info/features/featuredetails1186.html

Helderop, E., & Grubesic, T. (2021). Clustering Identification. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Information, B. S. of. (n.d.). Keeping Secrets: Anonymous Data Isn't Always Anonymous. Retrieved from https://ischoolonline.berkeley.edu/blog/anonymous-data

Knapp, E. (2021). Gentrification. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Missing Maps. (n.d.). One map myanmar and phandeeyar. Retrieved from https://www.missingmaps.org/blog/2018/09/24/mapping-in-myanmar/

O'Sullivan, D., & Manson, S. M. (2015). Do physicists have geography envy? And what can geographers learn from it? *Annals of the Association of American Geographers*, *105*(4), 704–722. Taylor & Francis.

QGIS.org (2021). QGIS Geographic Information System. *QGIS Association*. Retrieved from http://www.qgis.org.

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, *6*(3), 30–38.

Reades, J., & Smith, D. (2014). Mapping the “Space of Flows”: the geography of global business telecommunications and employment specialisation in the London Mega-City Region. *Regional Studies*, *48*(1), 105–126.

Reades, J., Zhong, C., Manley, E., Milton, R., & Batty, M. (2016). Finding pearls in london’s oysters. *Built Environment*, *42*(3), 365–381. Alexandrine Press.

Stephens, M. (2013). Gender and the geoweb: Divisions in the production of user-generated cartographic information. *GeoJournal*, *78*(6), 981–996. Springer.

Sweeney, L., Abu, A., & Winn, J. (2013). Identifying participants in the personal genome project by name (a re-identification experiment). Retrieved from http://arxiv.org/abs/1304.7605

Wolf, L. J. (2021). The Shape of Bias: Understanding the relationship between compactness and bias in US elections. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.

Xiao, N. (2016). *GIS algorithms: Theory and applications for geographic information science & technology*. Research methods. SAGE.

Ye, X., Peng, Q., & Gong, X. (2021). Network Analysis. In S. Rey & R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*. Edward Elgar.