Open Data for *Crime and Place* Research: A Practical Guide in R.

Samuel Langton (University of Leeds)

Reka Solymosi (University of Manchester)

**Keywords:** crime, open data, Open Street Map, transport, geography.

**Corresponding author:** Samuel Langton, University of Leeds, School of Law, Leeds, United Kingdom. Email: [s.langton@leeds.ac.uk](mailto:s.langton@leeds.ac.uk).

# Introduction

Access to data in crime and place research has traditionally been reserved for those who have the means to collect fresh data themselves, pay for access, or obtain data through formal data sharing agreements. Even when access is granted, the usage of these data often comes with conditions that circumscribe how the data can be used through licensing or policy (Kitchin, 2014). Even the public dissemination of findings which emerge from analysis might be subject to restrictions. This can lead to unequal access, controlled usage and curb the diffusion of findings, severely limiting the insight that can be obtained from data.

Open data initiatives provide a response to these shortcomings, broadening access and participation in research, removing the requirement for permissions, formal agreements, and negotiations (Manovich, 2011). Open data can lead to all sorts of novel insight within crime and place research, tapping into constructs and processes which are difficult to capture through surveys, interviews and other traditional measures (Solymosi & Bowers, 2018). As such, it is important that social scientists, researchers, crime analysts, and others interested in making sense of the world around them have the skills and know-how to access, interpret, critique, and analyse open datasets.

This chapter will outline how to develop such skills by providing a framework to approach and meaningfully interpret open data. The supplementary material also offers a practical hands-on guide to demonstrate how to access, wrangle, and analyse different sources of open data in order to draw conclusions about crime and place.

# Background

## What is open data?

First, it is useful to clarify what we mean when we refer to ‘open data’. The *Open Data Handbook*, compiled by the Open Knowledge Foundation, states that open data are “data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike” (Dietrich et al., 2009, para. 3). Specifically, open data can be defined along three key domains (Open Knowledge Foundation, 2016).

* **Availability and Access:** the data must be accessible via a public domain, at no more than a reasonable reproduction cost, and through a convenient medium, such as the internet. Ideally, it should be machine-readable and modifiable. For example, it should be downloadable from a website as an unlocked spreadsheet, or JSON file, rather than presented as a summary table in a PDF document.
* **Re-use and Redistribution:** the data must be provided under conditions that allow free use, such as modification, separation, or compilation of the data, and permit re-use and redistribution including the intermixing (merging) with other datasets.
* **Universal Participation:** everyone should be allowed to use, re-use and redistribute the data and its derivatives without restriction. For instance, restrictions that only allow non-commercial or educational usage would *not* constitute open data.

In short, ‘open data’ should be data that are readily and publicly available, in a usable format, and without restrictions on usage, modification and re-distribution.

Open data practices, and the extent to which it is a reality, vary both within and between countries. For example, the United States has a history of making public sector datasets openly available. The United Kingdom can be more strict, releasing data under a licence (‘Open Government Licence’), with some data requiring a fee (Kitchin, 2014). The *Global Open Data Index* (Open Knowledge Foundation, 2020) is one tool for tracking the annual global benchmark for publication of open government data by country. It provides a rank which indicates how well governments across the world follow open data practices across various domains. Many of these have direct relevance to crime and place research, such as administrative boundaries and geographic locations (e.g. coordinates).

The drive towards transparency and scrutiny of public information is further enhanced by an increasing need for replication. This is a key component of open science: transparent, reusable, reproducible and accessible research practices. Opening up datasets used in criminology publications and wider social sciences fosters and facilitates a culture of replication (Pridemore et al., 2018). As such, open data fits into the wider discourse around transparency, open review and open scrutiny, which (in time) may even become a policy requirement irrespective of the research field or the source of funding (Vuong, 2017). Increasingly, researchers are making use of online repositories for their papers (e.g. <https://arxiv.org/>), data (e.g. <https://www.ukdataservice.ac.uk/deposit-data>), code (e.g. <https://github.com/>), or whole projects (e.g. <https://osf.io/>) all in the name of transparency and replication. These are guidelines and tools worth considering, not just when conducting analysis on secondary data sources, but also collecting and sharing your own data.

## What are types of open data?

Now we have a reasonable idea about what open data is in a broad sense, we can consider the different types of open data which you might come across. Here, we formulate a typology of open datasets based on their origin.

### Public sector

The ‘public sector’ refers to organizations owned and operated by local or central government with the core aim of providing services to the public. Much of the open data movement has focused on opening up information generated by these local and national state agencies (Kitchin, 2014). These can include national statistics or administrative data, but also data collected by publicly funded research projects such as victimization and household surveys, amongst others. The *Global Open Data Index*, introduced earlier, focuses on identifying the gaps in governmental organizations, encouraging them to think about how public sector information can become more usable, and ultimately, more impactful (Open Knowledge Foundation, 2020).

Specific to crime and place research, datasets of interest in this particular domain might include data describing urban structure, such as street networks, to assess whether the configuration of roads dictates things like violent crime victimization (Summers & Johnson, 2017), or victimisation surveys to quantify citizens’ perceived safety and security.

### Private sector

In contrast to the public sector, ‘private sector’ tends to refer to parts of the economy which are not under direct state control, and often operate for profit. Opening up data generated by the private sector represents a significant challenge, largely due to the proprietary value to its creators (Kitchin, 2014). The aims and objectives of private companies, who have a duty to shareholders and operate in a competitive market environment, differ considerably from local and central government.

That said, some datasets are released openly by private sector organizations, albeit often only a subset of what would be available as a paying customer. The website for ArcGIS, a piece of software maintained by the Environmental Systems Research Institute (ESRI), a private company, host a number of open geographic datasets on their online cloud platform (<https://hub.arcgis.com/search>). Indeed, many papers have made use of data collected or distributed by private organizations to explore crime and place, such as Google Street View images (Langton & Steenbeek, 2017) and Twitter (Malleson & Andresen, 2015). In fact, you will learn how to obtain free Twitter data for such purposes in this book (see Chapter 6, Topic 7).

### Open crowdsourced data

Finally, there is a specific group of open data sources that collate data collectively, generated by large groups of individuals who do not specifically belong to any organization, but instead work together on a collaborative project. These are crowdsourced datasets. Examples include Wikipedia (<https://www.wikipedia.org/>), an online encyclopedia where anyone can contribute, or Flickr (<https://www.flickr.com/>) an online photo gallery where people can upload and tag their photos. Often, the organizations who maintain and monitor these data collection activities are charities or non-governmental organizations that do not operate for profit, but instead provide some form of social good. An example is the online reporting platform FixMyStreet (<https://www.fixmystreet.com/>), where people can report problems such as instances of graffiti, vandalism or environmental issues. This data has been used to explore signal crimes theory, for instance, offering insight into people’s experiences with incivilities (Solymosi et al., 2018). Another topic in this book (see Chapter 6, Topic 6) discusses the merits and pitfalls of crowdsourced data, and what to watch out for when analyzing open data of this type.

## Strengths and limitations

The defining characteristics of open data, namely, that of availability, re-usability and universal participation, outlined earlier, represent its greatest strength. But besides from this, there are other advantages. One specific motivation for using open data comes from its potential to address many of the limitations associated with traditional surveying methods, such as social desirability bias, or issues associated with memory and recall (Mayer-Schönberger & Cukier, 2013). This advantage exists largely due to the organic way in which open data is generated. It is often a by-product of other activities, and as such, we can gain an honest insight into people’s everyday lives and associated social processes (Solymosi & Bowers, 2018). As noted, open data also means open research, facilitating transparency and reproducibility. That said, there are a number of shortcomings which are worthy of consideration.

Firstly, one of the biggest obstacles to collecting open data is the ability to interpret what the data means (e.g. applying a framework for its analysis), and computational skills in data scraping, wrangling and cleaning, in order to transform it into a usable format for research (Boyd & Crawford, 2012). All sorts of open data remain inaccessible to people who may lack the skills and know-how to acquire them. Although this is certainly an obstacle for data usage more generally, it represents a key challenge to governmental bodies, in particular, on which there is an onus to ensure transparency and facilitate scrutiny. Private companies, especially those who generate open data as a by-product of their primary activity (e.g. Twitter), have little responsibility or pressure to ensure that their data is accessible and usable to the general public. Moreover, in both public and private spheres, licences and conditions on open data are not necessarily concrete, and might be subject to change with little or no notice. This is an important consideration when planning and running long-term research projects which involve open data.

Secondly, despite the merits of open data in terms of remedying pitfalls in traditional survey methods, there are other threats to validity that may emerge. For instance, whilst many online resources offer an ‘honest’ depiction of society, uncaptured by surveys, researchers should be careful in interpreting people’s communication online as completely authentic (Manovich, 2011). This could be a result of individuals willingly managing and ‘curating’ their online presence (Ellison et al., 2006), or because of wider issues such as government censorship, or cultural norms around particular topics, particularly sensitive topics of interest to researchers of crime, such as sexual assault or drug use.

Relatedly, researchers should be aware of issues over sampling, and ultimately, the generalizability of findings that emerge from the analysis of open data. In the case of crowdsourced data, the sample is not randomly drawn from a population, but rather, it is self-selected, giving way for people willing to discuss or contribute to a particular issue, which introduces a degree of bias (Longley, 2012). Specifically, contributors tend to be men, between the ages of 20-50, with a college or university degree (Budhathoki, 2010; Haklay, 2010). Contributions to resources such as Open Street Map (which we look at in the practical exercise) are correlated with contextual characteristics such as poverty and population density, and as such, coverage is non-uniformly distributed across urban areas (Mashhadi et al., 2013). These are important to keep in mind (and be transparent about) when reporting findings based on analysis of such data.

### What can be done?

Open data can be messy, biased and noisey, but criminology (and social sciences more generally) can benefit immeasurably from its use. Only through open data can public sector bodies be held to account, research be transparent and reproducible, and participation in data analysis universal. In crime and place research, both dependent variables (e.g. police-recorded crime incidents, victimization rates) and independent variables (e.g. demographic characteristics, ambient population estimates) can be sourced from open data, whether public, private or crowdsourced.

So, what can we do to make sure we make good use of these data? With a critical, engaged and considered approach to conducting research with open data, criminology can become a leading force in open and reproducible social science. In sum, researchers and analysts using open data must ask critical questions:

* **Where** does this data come from?
* **Why** was the data collected in the first place?
* **Who** is represented in the data, and who is excluded?
* **What** concepts and constructs can and cannot be operationalized with this data?
* **When** did data collection take place, and how might have that influenced results?

The answers to these questions will put the data sources in context, and ensure that researchers use them appropriately and with due care.

With this in mind, we encourage readers to complete the practical exercise in which we will utilize multiple sources of open data to explore police-recorded crime on and around public transport in London, England.

# Discussion

## Summary

This chapter has sought to introduce open data as a novel, growing and invaluable tool in crime and place research through a review of key fundamentals and a substantive demonstration within R. We have defined open data, and outlined some key types of open information available, along with their respective strengths and weaknesses. An important component of this review has been to encourage an engaged and critical research approach to using open data. Open data is both a rich resource of dependent and independent variables for researching the geography of crime, and an interesting research topic in and of itself as a tool for providing insight overlooked by traditional data sources. The supplementary material contained a practical exercise in which we accessed different sources of open data in order to explore crime concentrations at London Underground stations on the Jubilee line: police-recorded crime data via a direct download from a public sector website, transport data via directly querying a public sector API, hosted by Transport for London, and the equivalent transport data retrieved from a crowdsourced database, Open Street Map, using an API wrapper. Minor discrepancies in the geographic location of underground stations in each open data source painted two starkly different stories about the volume and concentration of crime occurring on the line. The demonstration highlighted both the power of open data and the need to critically engage with sources.

## Future of open data

As it stands, open data represents a key resource in crime and place research. It is a fundamental component in the movement towards transparent and reproducible scientific research. The skills required to effectively access and use open data, such as those demonstrated in this chapter, are increasingly being taught and deployed in social science. The number and quality of open data resources is also improving, many of which will help illuminate key topics in spatial criminology. For instance, information about building construction in Greater London is currently being collected through a crowdsourced platform, *Colouring London* (Hudson et al., 2019), opening prospect for studies to investigate historical urban development and crime concentrations in the capital, unrestricted by fees or strict terms, for the first time. But, as the discussions and demonstrations in this chapter have shown, there is still a long way to go. As with any emerging domain, open data sources vary considerably in their degree of completeness and reliability. Whilst this can offer insight, for instance, in terms of public participation and sense of community, it can also represent a significant obstacle to empirical analysis. A key challenge for the future will be to ensure that students and research practitioners ask critical questions of their data before blindly delving into analysis and public dissemination. Moreover, the longevity of open data licences remains an unknown parameter. With so much open data being distributed by the public sector, the accessibility of open information is subject to fluctuations in the amount of resource (and goodwill) available to safe-guard continuity. One way of justifying the continued collection and distribution of open public sector data is simply to use the data for public good, such as informing interventions which help reduce crime victimization.

# References

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Budhathoki, N. R. (2010). *Participants’ motivations to contribute geographic information in an online community* [PhD thesis]. University of Illinois at Urbana-Champaign.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, P., Tait, J., Zijlstra, T., & others. (2009). Open data handbook. In *Open Knowledge International*. <https://opendatahandbook.org/guide/en/what-is-open-data/>

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, *11*(2), 415–441.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of openstreetmap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703.

Hudson, P., Dennett, A., Russell, T., & Smith, D. (2019). Colouring london–a crowdsourcing platform for geospatial data related to london’s building stock. Proceedings from 27th annual gis research uk conference. *Newcastle University, Newcastle, 23-26 April.*

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE Publications.

Langton, S. H., & Steenbeek, W. (2017). Residential burglary target selection: An analysis at the property-level using google street view. *Applied Geography*, *86*, 292–299.

Longley, P. A. (2012). Geodemographics and the practices of geographic information science. *International Journal of Geographical Information Science*, *26*(12), 2227–2237.

Malleson, N., & Andresen, M. A. (2015). The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, *42*(2), 112–121.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the Digital Humanities*, *2*, 460–475.

Mashhadi, A., Quattrone, G., & Capra, L. (2013). Putting ubiquitous crowd-sourcing into context. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 611–622.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Open Knowledge Foundation. (2016). *Open definition version 2.1*. <https://opendefinition.org/od/2.1/en/>

Open Knowledge Foundation. (2020). *Global open data index*. <https://index.okfn.org/>

Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology*, *1*, 19–38.

Solymosi, R., & Bowers, K. (2018). The role of innovative data collection methods in advancing criminological understanding. *The Oxford Handbook of Environmental Criminology*, *210*, 210–237.

Solymosi, R., Bowers, K. J., & Fujiyama, T. (2018). Crowdsourcing subjective perceptions of neighbourhood disorder: Interpreting bias in open data. *The British Journal of Criminology*, *58*(4), 944–967.

Summers, L., & Johnson, S. D. (2017). Does the configuration of the street network influence where outdoor serious violence takes place? Using space syntax to test crime pattern theory. *Journal of Quantitative Criminology*, *33*(2), 397–420.

Vuong, Q. H. (2017). Open data, open review and open dialogue in making social sciences plausible. *Nature: Scientific Data Updates*. <https://ssrn.com/abstract=3086667>