# Open Data

Samuel Langton (MMU) & Reka Solymosi (University of Manchester)

June 2020

## 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 and 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 data sets.

This chapter will outline how to develop such skills by providing a framework to approach and meaningfully interpret open data. The chapter 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). Specifically, open data can be defined along three key domains (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 data sets 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* (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 - a hallmark for open science research practice. Opening up data sets used in criminology publications and wider social sciences fosters and facilitates a culture of replication (Pridemore, Makel, and Plucker 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://osf.io/), data (e.g. https://www.ukdataservice.ac.uk/deposit-data) and code (e.g. https://github.com/), 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 data sets based on their origin.

### Public sector

The 'public sector' refers to organizations owned and operated by local or federal government with the core aim to provide 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 think about how public sector information can become more usable, and ultimately, more impactful (Foundation 2020).

Specific to crime and place research, data sets 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 victimisation (Summers and 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 and Steenbeek 2017) and Twitter (Malleson and Andresen 2015). In fact, you will learn to 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 data sets. 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, Bowers, and Fujiyama 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 analysing 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 and 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 and 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 and 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, Heino, and Gibbs 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 later) are correlated with contextual characteristics such as poverty and population density, and as such, coverage is non-uniformly distributed across urban areas (Mashhadi, Quattrone, and Capra 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 first 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 a context of understanding, and ensure that the researchers use them appropriately and with care.

With this in mind, we now move on to a 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.

# Practical exercise

In this exercise you will acquire a number of different skills, including:

- Accessing open data using three different methods:
  - Direct download.
  - Direct calls to an Application Programming Interface (API).
  - Calls to an API using a wrapper.
- Cleaning, wrangling, and visualizing open spatial data.
- Comparing different sources of open data.
- How to engage with the critical 'Where', 'Why', 'Who', 'What' and 'When' questions to better understand open data.

## Our aim

In this exercise we will explore crime in and around London Underground stations. We know that environmental features are important when it comes to public transport areas being more or less criminogenic. Studies in Sweden, for instance, have shown the importance of environmental and neighborhood characteristics in determining crime concentrations at underground stations in Stockholm, along with the positioning of stations on the line (Ceccato, Uittenbogaard, and Bamzar 2013). Similar work has been carried out exploring bus stops in Los Angeles (Loukaitou-Sideris 1999) and the impact of intensive policing along bus corridors in Merseyside, England (Newton, Johnson, and Bowers 2004), amongst others.

Here, we will consider the case of London. Specifically, we will examine the question: to what extent does crime cluster in and around London Underground stations? We will use various sources of open data to answer this question. This will allow us to explore the strengths and limitations of public sector and crowdsourced open data sources.

## Accessing data

We will be using two different types of open data in this exercise: public sector data (police recorded crime data and local transport authority data) and crowdsourced data (from Open Street Map). To access them, we will use three different methods, namely, (1) direct download, (2) direct request to an API, 3) request to an API using a wrapper. This will give you a few different ideas about how you might go about accessing other open data sets relevant to your research. Locating, identifying and learning to access open data is a skill in itself.

### Direct download

The simplest way that open data can be made available is through direct download from a website. In such a case, you can visit a website, select some parameters, and save a file containing the data you requested locally on your computer.

In the United Kingdom, police recorded crime data in England and Wales can be accessed this way using an online web portal (https://data.police.uk/) under Open Government Licence. Visit this website, and you will see a welcome message, and six tabs across the top of it, which should read "Home", "Data", "API", "Changelog", "Contact", and "About".

Before we download any data, we can learn more about it by clicking on the "About" tab. This brings up a lot of information that it is important to review carefully in order to answer the where, why, who, what and when questions posed earlier. Take a moment to read through this information, and try to take notes on what you think might be relevant for your analysis. For example, if we want to map crimes, we might want to explore if there is any type of "anonymization" that might take place before the data are released (to protect the privacy of the victims). If you read the "About" page, you might find the following note:

> Location anonymization The latitude and longitude locations of Crime and ASB incidents published on this site always represent the approximate location of a crime — not the exact place that it happened.

This indicates that although we get a latitude and longitude coordinate with each crime event, it may only be approximate/ This may have implications for our findings later!

To then actually download some data, move on to clicking on the "Data" tab. This should open a page entitled "Data downloads" under which you can see another five tabs: "Custom download", "Archive", "Boundaries", "Open data", and "Statistical data". By staying on this "Custom download" page you can select what sort of data you want to download. We can select the time period and police force of interest, and the type of information required (e.g. crimes, stop and search, outcomes). We are also informed that the data are downloaded in comma-separated values format (.csv file extension), which meets our machine-readable, easy-to-manipulate data format requirements.

For this exercise, we are going to use British Transport Police data, a force which operates on railways and light-rail systems across the country, for the month of January in 2020. Select the time period using the dropdown menus, the force using the tickboxes, and then generate and download the file.

Save this file locally (in your working directory) to a subfolder named "data". You can load it in with the `read_csv()` command from the `readr` package. Remember that you will need to load it with `library(readr)`.

```r
library(readr)
btp_df <- read_csv("data/2020-01-btp-street.csv")
```

```
## Parsed with column specification:
## cols(
##   `Crime ID` = col_logical(),
##   Month = col_character(),
##   `Reported by` = col_character(),
##   `Falls within` = col_character(),
##   Longitude = col_double(),
##   Latitude = col_double(),
##   Location = col_character(),
##   `LSOA code` = col_character(),
##   `LSOA name` = col_character(),
##   `Crime type` = col_character(),
##   `Last outcome category` = col_logical(),
##   Context = col_logical()
## )
```

Note: If you didn't manage to follow along with the download instructions, you can also access this dataset from our github page. Note that the line break below is purely for easy formatting - this will need removing in your own script so that the URL runs across one line.

```
btp_df <- read_csv("https://raw.githubusercontent.com/langtonhugh/osm_crim/master/data/
                   2019-01-greater-manchester-street.csv")
```

You should now have a data frame with 5694 crimes in, ready for you to explore!

### Direct request to an API

Another way of downloading open data is through an Application Programming Interface (API). This is a tool which defines an interface for a programme to interact with a software component. For example, it defines the sort of requests or calls which can be made, and how these calls and requests can be carried out. Here, we are using the term 'API' to denote tools created by an open data provider to give access to different subsets of their content. Such APIs facilitate scripted and programmatic extraction of content, as permitted by the API provider (Olmedilla, Martínez-Torres, and Toral 2016). APIs can take many different forms and be of varying quality and usefulness (Foster et al. 2016). For the purposes of accessing open data from the web, we are specifically talking about *RESTful* APIs. The 'REST' stands for Representational State Transfer. These APIs work directly over the web, which means users can play with the API with relative ease in order to understand how it works (Foster et al. 2016).

Here, we will make a direct request to the API created by London's local government organization responsible for transport: Transport for London (TfL). TfL oversee the London Underground. They provide access to their open data through a unified API. Much like how we found the "About" page when downloading police data directly, to answer our 'Where', 'Why', 'Who', 'What', 'When' questions, we must find a similar document for TfL. Details about the TfL API are provided through their open data page (https://tfl.gov.uk/info-for/open-data-users/unified-api), which explains what data is available, and how the API is designed.

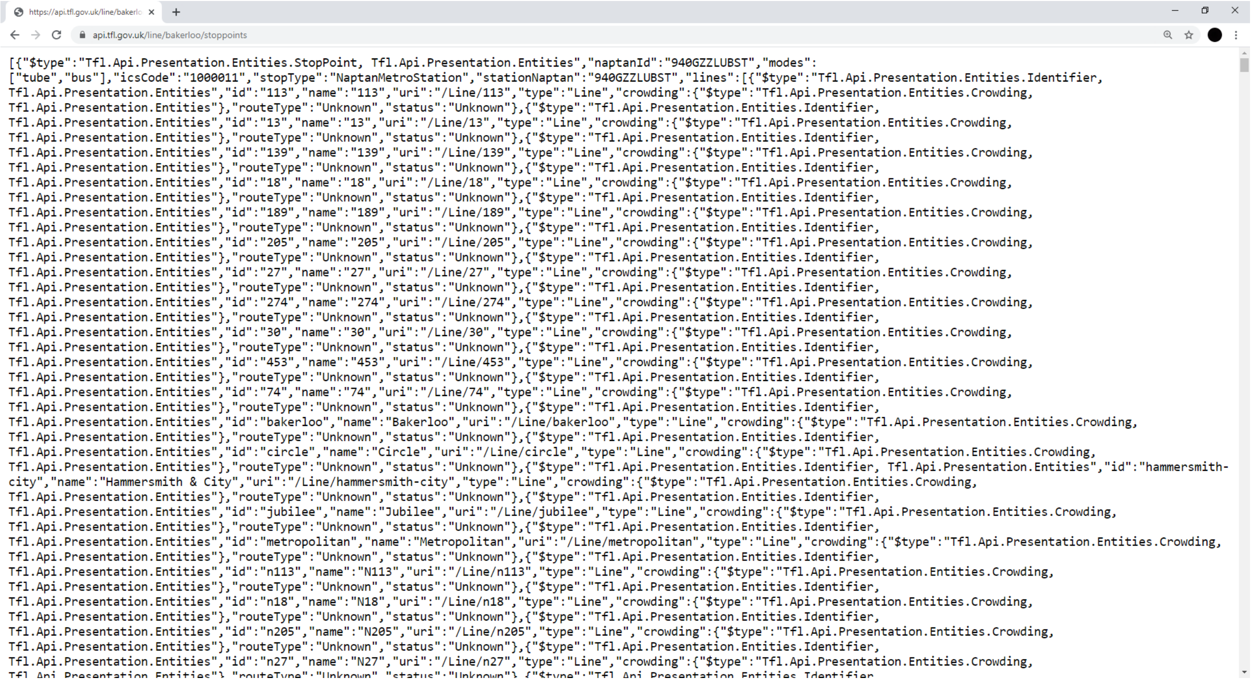
They further provide a documentation page with examples of how you can use HTTP (i.e. web link) requests to make calls to the API (see: https://api-portal.tfl.gov.uk/docs). If you follow this link you should see a page titled "Our Unified API" and some examples of calls.

To draw conclusions about crimes in and around London Underground stations, we will need some spatial data about the stations themselves.

In the documentation, there is a section called 'API area' which offer some guidance on the types of data available. Scroll down and find the example called *Stops*. You will see that this call returns information on stops for buses or London Underground lines, and there are two examples (bus route 24 and the Bakerloo line). Copy their demonstration URL for the bakerloo line (https://api.tfl.gov.uk/line/bakerloo/stoppoints) and paste it into your web browser. When you visit this page you should see something like Figure 1.

Although this contains the information we requested, it is not very legible for our human eyes. It is in a format called *JSON*. This stands for JavaScript Object Notation. JSON is an open standard way of storing and exchanging data, and will most likely be the format in which data are returned from most API calls. As you can see, it is not intuitive to read in its current format *but* it is machine-readable friendly, which again aligns well with our requirements for open data.

In reality, you won't usually make these calls by pasting them into a web browser, but it is useful to see how it works. More likely, we will actually make these calls from within the R environment. That said, it


https://api.tfl.gov.uk/line/bakerlo
api.tfl.gov.uk/line/bakerloo/stoppoints

```
[{"$type":"Tfl.Api.Presentation.Entities.StopPoint, Tfl.Api.Presentation.Entities","naptanId":"940GZZLUBST","modes":
["tube","bus"],"icsCode":"1000011","stopType":"NaptanMetroStation","stationNaptan":"940GZZLUBST","lines":[{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"113","name":"113","uri":"/Line/113","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"13","name":"13","uri":"/Line/13","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"139","name":"139","uri":"/Line/139","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"18","name":"18","uri":"/Line/18","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"189","name":"189","uri":"/Line/189","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"205","name":"205","uri":"/Line/205","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"27","name":"27","uri":"/Line/27","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"274","name":"274","uri":"/Line/274","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"30","name":"30","uri":"/Line/30","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"453","name":"453","uri":"/Line/453","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"74","name":"74","uri":"/Line/74","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"bakerloo","name":"Bakerloo","uri":"/Line/bakerloo","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"circle","name":"Circle","uri":"/Line/circle","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier, Tfl.Api.Presentation.Entities","id":"hammersmith-
city","name":"Hammersmith & City","uri":"/Line/hammersmith-city","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"jubilee","name":"Jubilee","uri":"/Line/jubilee","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"metropolitan","name":"Metropolitan","uri":"/Line/metropolitan","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"n113","name":"N113","uri":"/Line/n113","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"n18","name":"N18","uri":"/Line/n18","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"n205","name":"N205","uri":"/Line/n205","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
Tfl.Api.Presentation.Entities","id":"n27","name":"N27","uri":"/Line/n27","type":"Line","crowding":{"$type":"Tfl.Api.Presentation.Entities.Crowding,
Tfl.Api.Presentation.Entities"},"routeType":"Unknown","status":"Unknown"},{"$type":"Tfl.Api.Presentation.Entities.Identifier,
```

Figure 1: example of data from Transport for London

might be useful (and interesting) to examine how basic queries are constructed using this documentation. For example, we now know that to request the stops from the "Bakerloo" line, we need the URL:

`"https://api.tfl.gov.uk/line/bakerloo/stoppoints"`

Let's say that instead, we want data for the Northern Line. What do you think that URL will look like? Indeed, all you need to do is replace "bakerloo" with "northern".

In R, we can use the `fromJSON()` function from the `jsonlite` package to parse all this information into a data frame (with rows and columns). It will then become more familiar and usable. All we need to do is input the URL from the TfL API into the function with a bit of help from `readLines()`, a function in base R for reading text from URLs.

To keep things focused, let's request data about stop points on the Northern line, one of the largest and busiest on the network. Note how this is simply an amended version of the example provided by TfL in the API documentation.

```r
library(jsonlite)

api_call <- fromJSON(readLines("https://api.tfl.gov.uk/line/northern/stoppoints"))
```

```
## Warning in readLines("https://api.tfl.gov.uk/line/northern/stoppoints"):
## incomplete final line found on 'https://api.tfl.gov.uk/line/northern/stoppoints'
```

This gives us an object (`api_call`) which contains all the information returned by the TfL API. JSON is slightly different to traditional data frames with rows and columns, which we are probably more familiar with, because the data is nested. For instance, `api_call` is classed as a data frame, but now try viewing the object using `View(api_call)`. You will notice that some of the columns are actually lists, rather than character or factor vectors, which what we might usually expect. Another way of exploring the `api_call` object is by looping the `class()` function through all the columns in the data frame using `lapply(api_call, class)`.

This demonstrates an important challenge faced by researchers when using open data, because dealing with data in this format can be messy and complicated. It is not always a neatly formatted data frame like the csv from the open police data portal. That said, we can transform this data into something more familiar from within R, as we will see later on.

### Request to an API using a wrapper

Finally, we will look at how open data can be obtained using API wrappers. Often, developers who work with APIs will share their code, and release them in the form of a package or module, so that other people can use it. This is called a *wrapper* because it uses code that 'wraps' around the API to make it a neater, more usable package. Wrappers remove (or at least lower) many of the obstacles to accessing open data noted earlier. The wrapper can take many forms, such as a Python module, or an R package. It could even be a web interface that provides a graphical user interface (GUI) for accessing the API in question.

To demonstrate this, we will be accessing data from Open Street Map, a database of geospatial information built by a community of mappers, enthusiasts and members of the public, who contribute and maintain data about all sorts of environmental features, such as roads, green spaces, restaurants and railway stations, amongst many other things, all over the world. As such, it is a prime example of 'crowdsourced' open data. You can view the information contributed to Open Street Map using their online mapping platform (https://www.openstreetmap.org/). The result of people's contributions is a database of spatial information rich in local knowledge which provides invaluable information about places and their features, without being subject to strict terms on usage.

Open Street Map has two types of wrappers available for its API, a web-based GUI called Overpass Turbo (https://overpass-turbo.eu/), and an R package called `osmdata`.

If we load the package `osmdata` we can use its functions to query the Open Street Map API, rather than the API query being made directly (like we did for TfL, using the URL). Once again, we will want to identify documentation which can help us understand and critique our data, and learn how to query it. When it comes to OSM, you can read all about it on their community page: https://www.openstreetmap.org/about. Using the wrapper in R, we can refer to the package documentation and the associated vignette which we can find at https://cran.r-project.org/web/packages/osmdata/vignettes/osmdata.html.

Unlike the TfL API, `osmdata` is an international database, and has lots of data that we might not necessarily need. As such, the first thing we need to specify is our study region. To do this, we use the `getbb()` function from the `osmdata` package, which stands for "get bounding box". You can think of the bounding box as a box drawn around the area that we are interested in (in this case, London, England) which tells the OSM API that we want everything *inside* the box, but nothing *outside* the box.

So, how can we name a bounding box specification to define the study region? This can be obtained manually, which requires some existing knowledge about an area using the latitude and longitude coordinates, or you can use a search term. Here, we want to select Greater London, so we can use the search term "greater london united kingdom". Besides specifying the study region, we can also tell the `getbb()` function what format we want the data to be in. In this case we want a spatial object, specifically an `sf` polygon object, which we name `bb_sf`:

```r
library(osmdata)
```

```
## Data (c) OpenStreetMap contributors, ODbL 1.0. https://www.openstreetmap.org/copyright
```

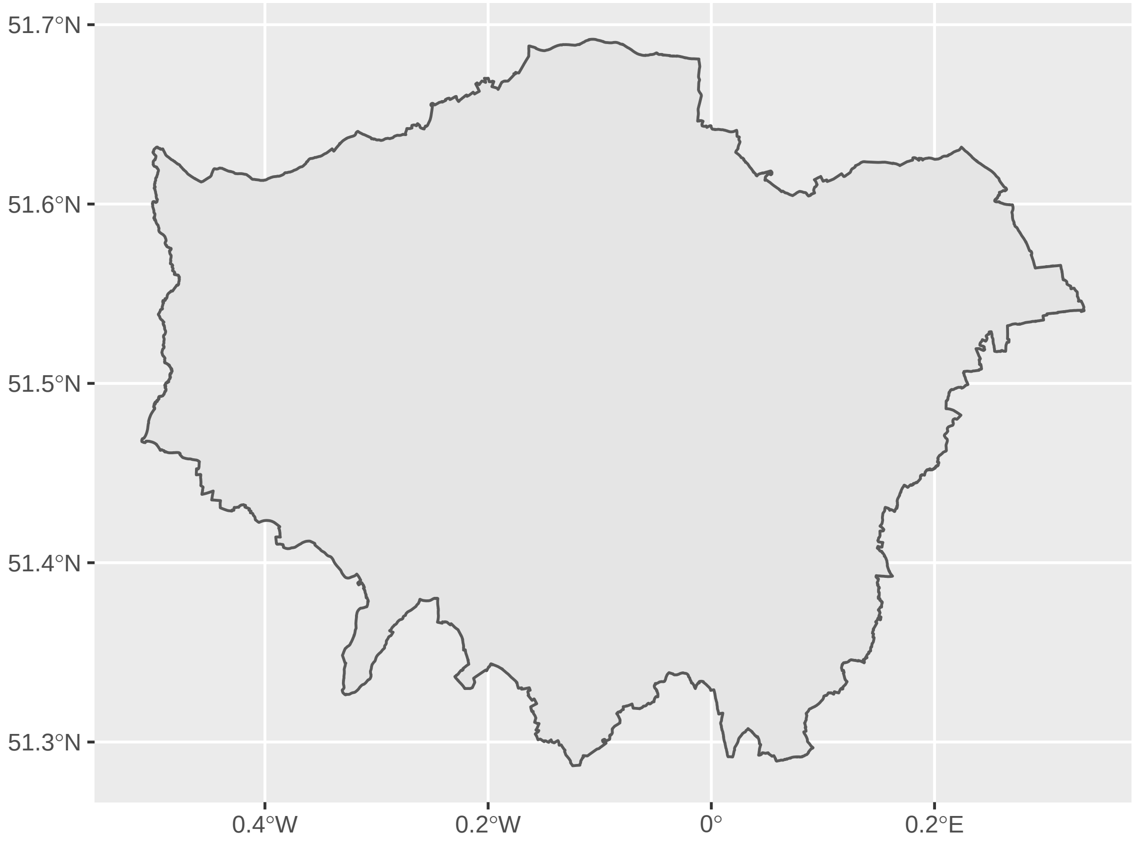
```r
bb_sf <- getbb(place_name = "greater london united kingdom", format_out = "sf_polygon")
```

We now have our study region defined as the administrative boundaries of Greater London. We can visualize this quickly to check that it has worked as expected using the `sf` and `ggplot2` packages.

```r
library(ggplot2)
library(sf)
```

```
## Linking to GEOS 3.6.1, GDAL 2.2.3, PROJ 4.9.3
```

```r
ggplot(data = bb_sf) +
  geom_sf()
```



If you are familiar with the shape of London then this image should be quite recognizable! We can now move on to scrape data from the Open Street Map API using the `opq()` function. The function name is short for 'Overpass query', which is how users can query the Open Street Map API using search criteria.

Besides specifying what area we want to query with our bounding box object (`bb_sf`) in the `opq()` function, we must also define the feature which we want to pull from the API. Features in Open Street Map are defined through 'keys' and 'values'. Keys are used to describe a broad category of features (e.g. highway, amenity), and values are more specific descriptions (e.g. cycleway, bar). These are tags which contributors to Open Street Map have defined. A useful way to explore these is by using the comprehensive Open Street Map Wiki page on map features (https://wiki.openstreetmap.org/wiki/Map_Features).

We can select what features we want using the `add_osm_feature()` function, specifying our key as 'public transport' and our value as 'station'. We also want to specify what sort of object (what class) to get our data into, and as we're still working with spatial data, we stick to the `sf` format, for which the function is `osmdata_sf()`.

Finally, we trim everything down to our study region with the `trim_osmdata()` function. Without this addition, we would get points for everywhere within our bounding box, including some areas outside of Greater London. We add all these together using a pipe (`%>%`) and make the call like so:

```r
osm_stat_sf <- opq(bbox = bb_sf) %>%                                    # select bounding box
  add_osm_feature(key = 'public_transport', value = 'station') %>% # select features
  osmdata_sf() %>%                                                  # specify class
  trim_osmdata(bb_poly = bb_sf)                                     # trim by bounding box
```

The resulting object `osm_stat_sf` contains lots of information. We can view the contents of the object by simply executing the object name into the **Console**.

```
osm_stat_sf
```

```
## Object of class 'osmdata' with:
##                  $bbox : 51.2867601,-0.5103751,51.6918741,0.3340155
##         $overpass_call : The call submitted to the overpass API
##                  $meta : metadata including timestamp and version numbers
##            $osm_points : 'sf' Simple Features Collection with 1380 points
##             $osm_lines : NULL
##          $osm_polygons : 'sf' Simple Features Collection with 44 polygons
##        $osm_multilines : NULL
##     $osm_multipolygons : 'sf' Simple Features Collection with 1 multipolygons
```

This confirms details like the bounding box coordinates, but also provides information on the features collected from the query. As one might expect, most information relating to public transport station locations has been recorded using points (i.e. two-dimensional vertices, coordinates) of which we have over seven thousand at the time of writing. We also have around one hundred polygons. For now, let's extract the point information.

```
osm_stat_sf <- osm_stat_sf$osm_points
```

We now have an `sf` object with all the public transport stations in London mapped by OSM volunteers, along with 130 variables of auxiliary data, such as the "fare_zone" the station is in, what "amenity" it may have and whether it has "toilets", amongst many others. Of course, it is up to the volunteers whether they collect all these data, and in many cases, they have not have added information. Nevertheless, when the details are recorded, they provide rich insight and local knowledge we may otherwise not be able to obtain.

One additional step needed is to select the stations which are relevant to us (i.e. they fall along the Northern Line). The variable `line` can help us with this. First, let's look at the values in this variable by using the `unique()` function:

```
unique(osm_stat_sf$line)
```

```
##  [1] NA
##  [2] "Northern"
##  [3] "District;Circle"
##  [4] "Metropolitan"
##  [5] "Piccadilly"
##  [6] "Victoria"
##  [7] "Northern City"
##  [8] "Central"
##  [9] "District"
## [10] "Circle;District"
## [11] "Bakerloo"
## [12] "District, Piccadilly"
## [13] "Metropolitan;Jubilee"
## [14] "Metropolitan, Piccadilly"
## [15] "Jubilee"
## [16] "District, Circle, Piccadilly"
## [17] "Metropolitan;Circle;Hammersmith & City"
## [18] "Overground;Victoria"
## [19] "Overground"
```

You might see that the Northern line appears a few times, once as "Northern" but also once as "Northern City". We know that there is no line called "Northern City", and we can see that many other combinations of these values exist (e.g. "Circle;District" indicates that the station serves both Circle and District lines).

Therefore, we can assume that stations like “Northern City” serve both the Northern line and one other. To get round this, and select all stations with “Northern” somewhere in the name, we can use the `grepl()` function from base R. Note that we also make use of `filter()` from within `dplyr`, so we first need to load that package.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
osm_north_sf <- osm_stat_sf %>%
  filter(grepl("Northern", line))
```

Now we have all our open data sets loaded into the R environment!

## Cleaning, wrangling, and visualising our data

Obtaining open data from a direct download or API is often only half the battle. It doesn’t necessarily mean that the data is in the shape needed to conduct analysis. The first thing to do once we have acquired open data is investigate it using data cleaning, wrangling, and visualisation. We’ve already loaded the `dplyr`, `sf` and `ggplot2` packages, but here, we will also make use of `tidyr` and `patchwork`, so load these now. Remember, if you do not have these packages installed, use the `install.packages()` function prior to loading each one with `library()`.

```
library(tidyr)
library(patchwork)
```

### From JSON to data frame

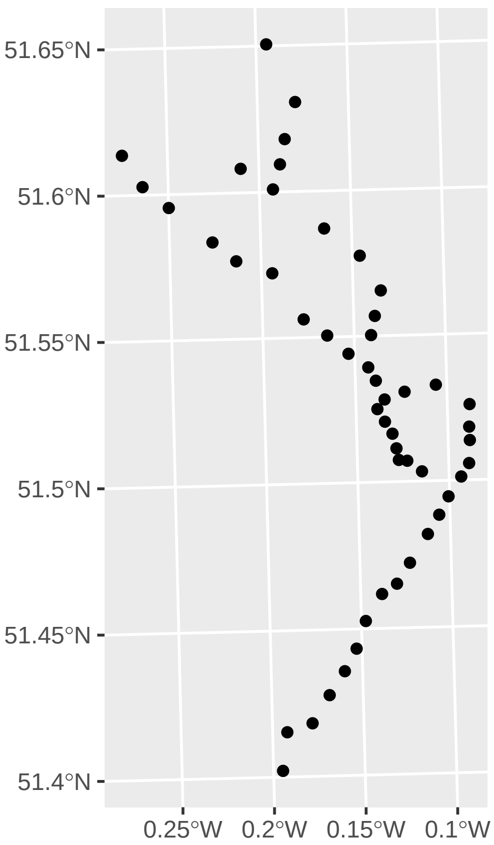
We usually want to work with data that is in the format of a data frame, that is, all your observations are rows and all your variables are columns. While this is the case for our police data and our OSM data, recall that the TfL data came in the nested JSON format.

Fortunately for us, R is more than capable of dealing with nested data. Moreover, we can see that the information we are interested in (namely, the location of underground stations) has already been successfully parsed into columns called **lon** (longitude) and **lat** (latitude), along with an identification column of the station names called **commonName**. We can extract these columns, and create an `sf` (spatial) point object using these coordinates, all in one chunk of code, using the piping operator (`%>%`). Note that we define the Coordinate Reference System (CRS) which the data has come in (WGS 84), and then transform it into something more appropriate (the British National Grid, a projected CRS used in the United Kingdom).

```
tfl_north_sf <- api_call %>%
  select(commonName, lat, lon) %>%
  st_as_sf(coords = c(x = "lon", y = "lat"), crs = 4326) %>%
  st_transform(27700)
```

We can now easily visualize this data using the `ggplot2` package, which is compatible with `sf` objects.

```
ggplot(tfl_north_sf) +
  geom_sf()
```

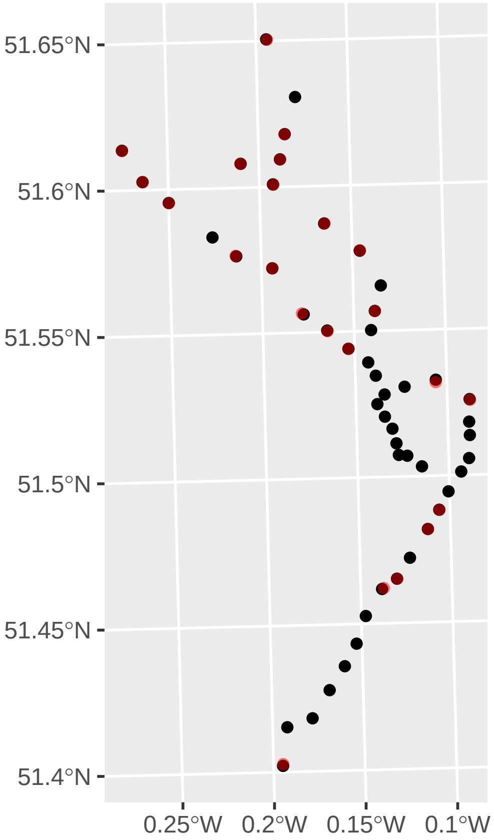


There we have it: with just a few lines of code in R, we have queried the TfL API, a public sector source of open data, and plotted a basic map of stations on the Northern underground line.

### Visualise to check and compare our data sets

We can now plot the OSM station points, which are already spatial, over these station locations pulled from the TfL API, to see how the two data sets compare. We do not plot the study region boundary in order to get a more detailed, local view.

```r
ggplot() +
  geom_sf(data = tfl_north_sf) +
  geom_sf(data = osm_north_sf, color = "red", alpha = 0.5)
```



As we can see, there is clearly some discrepancy between the two data sources. There are slight differences in the exact point locations of stations, but more saliently, Open Street Map appears to be missing stations, especially in Central and South London. This could be due to a number of reasons, among them: variation in residents' willingness to contribute to information about their local area, and the demographic characteristics known to influence the volume of contributions (Mashhadi, Quattrone, and Capra 2013). Given that South London is historically less wealthy in comparison to the North, for instance, this would certainly be an interesting avenue to pursue. Interestingly, these missing stations *have* actually been recorded by contributors in Open Street Map, but *without* information about the line on which the station is located. We can see this by plotting out all Open Street Map stations using `osm_stat_sf`. That said, without local knowledge, an alternative open data source, or a critical approach to using open data, well-intentioned researchers might draw erroneous conclusions from filtering stations tagged as on the Northern line. This gives us a preliminary answer to our earlier questions: the public sector open data from TfL appears to hold some advantage over its crowdsourced alternative, Open Street Map, in terms of London Underground lines.

But, what might the impact of this be on studying crime in and around London Underground stations, given the differences observed between the two open data sources? For that, we can turn to our British Transport Police data.

In its raw form, this data is a standard data frame, but we can make it spatial using some of the functions we used earlier. Note that we need to add `st_intersection()` to clip our points to our study region, because the British Transport Police are operational across the country, not just in London. To do this, we first project `bb_sf` to the British National Grid, so that we perform the intersection on data which have the same projection.

```
# First, project the study region boundaries.
bb_sf <- st_transform(bb_sf, 27700)

# Then, make the police data spatial, and clip.
```

```r
btp_sf <- btp_df %>%
  drop_na(Longitude, Latitude) %>%
  st_as_sf(coords = c(x = "Longitude", y = "Latitude"), crs = 4326) %>%
  st_transform(27700) %>%
  st_intersection(bb_sf)
```

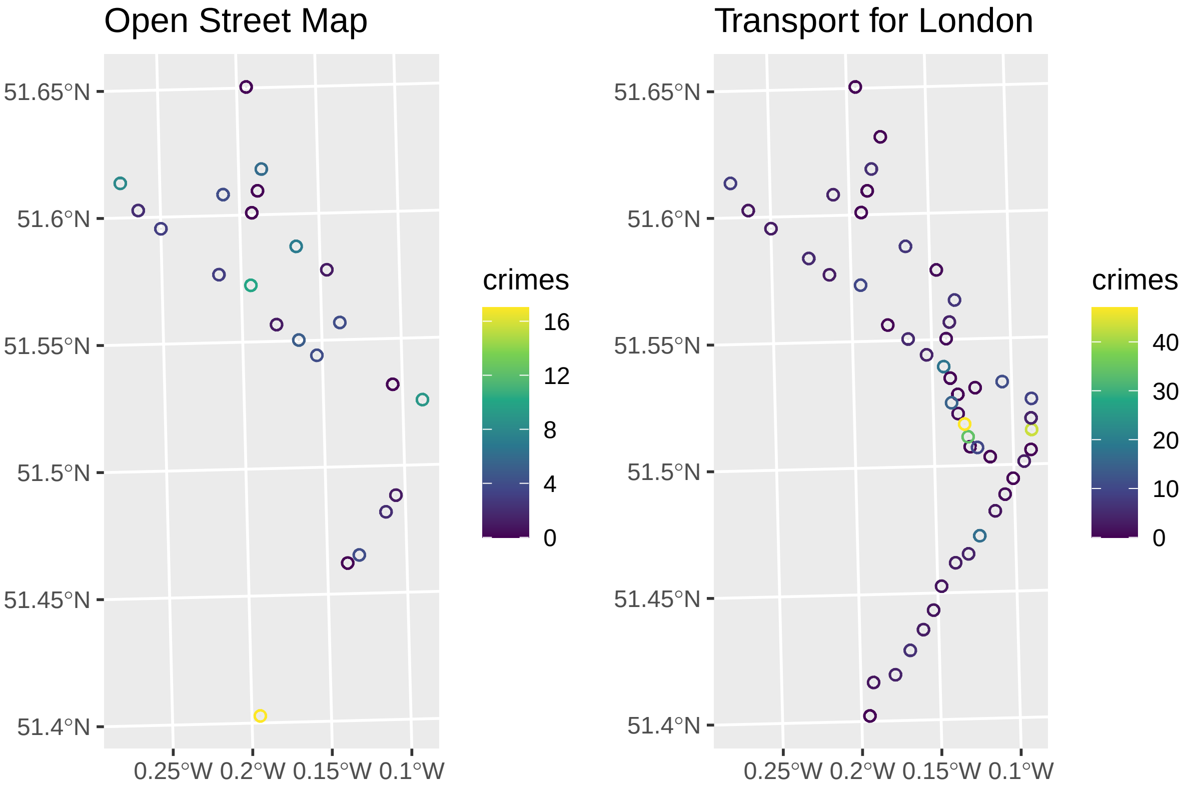
```
## Warning: attribute variables are assumed to be spatially constant throughout all
## geometries
```

We now have the location of all crimes recorded by the British Transport Police in Greater London during January, 2020. For the purposes of this demonstration, we can define crimes occurring as ‘in and around’ Northern line tube stations by creating a 50 meter buffer around each station, for each source of data, and counting the number of points falling within each. It is worth noting that this definition is somewhat arbitrary, and is subject to the spatial inaccuracy in open police recorded crime data [see tompson2015uk], but it does serve to facilitate this demonstration.

```r
# Assign the same CRS.
osm_north_sf <- st_transform(osm_north_sf, 27700)
tfl_north_sf <- st_transform(tfl_north_sf, 27700)

# Create buffers to define 'in and around'.
osm_buff_sf <- st_buffer(osm_north_sf, dist = 50)
tfl_buff_sf <- st_buffer(tfl_north_sf, dist = 50)

# Count number of crimes recorded within each buffer.
osm_north_sf <- osm_buff_sf %>%
  mutate(crimes = lengths(st_intersects(osm_buff_sf, btp_sf)))
tfl_north_sf <- tfl_buff_sf %>%
  mutate(crimes = lengths(st_intersects(tfl_buff_sf, btp_sf)))
```

We can then visualize these counts and compare the two sources of data by coloring in our point locations according to the crime count. Note that we arrange the plots side-by-side using syntax available using the `patchwork` package.

```r
library(patchwork)

p1 <- ggplot(data = osm_north_sf) +
  geom_sf(mapping = aes(color = crimes), size = 2) +
  labs(title = "Open Street Map") +
  scale_color_viridis_c()

p2 <- ggplot(data = tfl_north_sf) +
  geom_sf(mapping = aes(color = crimes), size = 2) +
  labs(title = "Transport for London") +
  scale_color_viridis_c()

p1 + p2
```



A number of factors emerge from this visualization. First, issues arising from the disparity between the stations identified on the Northern line are compounded. Using the Open Street Map data in isolation, we are grossly underestimating the number of crimes occurring in and around Northern line stations. This could have significant implications for how police allocate resource across the underground network. Secondly, as a result of this, the spatial patterning of crime on underground stations differs considerably between the two data sources. Using the Open Street Map data, the final station on the line, Morden, has the greatest concentration of crimes, whereas the TfL data suggests that the most problematic station is Tottenham Court Road, in the city centre. The conclusions drawn are starkly different, and would determine the consistency of findings with existing research examining crime at underground stations (Ceccato, Uittenbogaard, and Bamzar 2013).

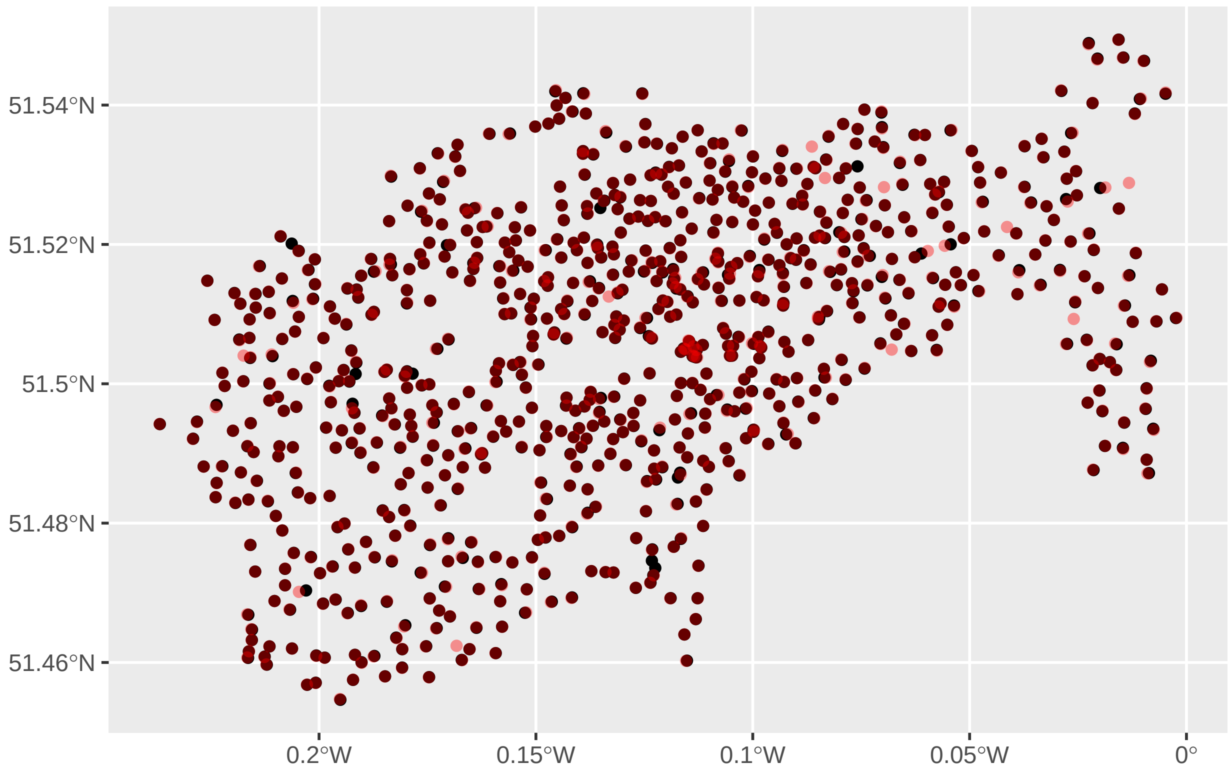
That said, this is not to say that Open Street Map is an inadequate source of open data for crime and place research. Firstly, it is an ever-growing resource. It is plausible (indeed, likely) that in the near future, the missing information about London Underground lines will be completed. Secondly, as noted, systematic variation in the extent to which individuals have contributed to the database can generate interesting discussions and avenues of research in its own right. Thirdly, this is simply one of thousands of features contained in the Open Street Map database, many of which are more complete. If we were interested in examining concentrations of bicycle theft from bike rental docking stations (so-called “Boris bikes”) in Greater London, we could query the APIs from TfL and Open Street Map, just as we did earlier, but using slightly amended criteria. This is demonstrated below, all in one code chunk.

```
bb_sf <- getbb(place_name = "greater london", format_out = "sf_polygon")

osm_bike_sf <- opq(bbox = bb_sf) %>%
  add_osm_feature(key = 'amenity', value = 'bicycle_rental') %>%
  osmdata_sf() %>%
  trim_osmdata(bb_poly = bb_sf)
```

```
osm_bike_sf <- osm_bike_sf$osm_points

osm_bike_sf <- osm_bike_sf %>%
  filter(brand == "Santander Cycles")

tfl_bike_sf <- fromJSON(readLines("https://api.tfl.gov.uk/bikepoint"))
```

```
## Warning in readLines("https://api.tfl.gov.uk/bikepoint"): incomplete final line
## found on 'https://api.tfl.gov.uk/bikepoint'
```

```
tfl_bike_sf <- tfl_bike_sf %>%
  st_as_sf(coords = c(x = "lon", y = "lat"), crs = 4326)

ggplot() +
  geom_sf(data = osm_bike_sf) +
  geom_sf(data = tfl_bike_sf, col = "red", alpha = 0.4)
```



As showcased, the resulting visualizations are highly similar (a Spatial Point Pattern Test will confirm this, see Chapter 8, Topic 4). Refer back to the example earlier if some of the code is unclear. Note that this is a major benefit of using an API wrapper from within R, because the code used to query each open data source can easily be re-used, with minor amendments, to extract different features.

# 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. By way of an example, we accessed different sources of open data in order to explore crime concentrations at London Underground stations: 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. We found key differences between the two sources of transport data. Specifically, missing information in Open Street Map resulted in a skewed picture of crime occurring in and around underground stations. This highlighted both the power of open data, but also its shortcomings. We concluded with an additional demonstration using bike rental docking stations in Greater London, which showcased a scenario in which both public sector and crowdsourced open data provide unique (and comparable) insight.

## 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 which traces building construction in Greater London is currently being collected through a crowdsourced platform, *Colouring London* (Hudson et al. 2018), 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, Danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–79.

Budhathoki, Nama R. 2010. "Participants' Motivations to Contribute Geographic Information in an Online Community." PhD thesis, University of Illinois at Urbana-Champaign.

Ceccato, Vania, Adriaan Uittenbogaard, and Roya Bamzar. 2013. "Security in Stockholm's Underground Stations: The Importance of Environmental Attributes and Context." *Security Journal* 26 (1): 33–59.

Dietrich, Daniel, Jonathan Gray, Tim McNamara, Antti Poikola, P Pollock, Julian Tait, Ton Zijlstra, and others. 2009. "Open Data Handbook." *Open Knowledge International*.

Ellison, Nicole, Rebecca Heino, and Jennifer Gibbs. 2006. "Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment." *Journal of Computer-Mediated Communication* 11 (2): 415–41.

Foster, Ian, Rayid Ghani, Ron S Jarmin, Frauke Kreuter, and Julia Lane. 2016. *Big Data and Social Science: A Practical Guide to Methods and Tools*. crc Press.

Foundation, Open Knowledge. 2016. "Open Definition 2.1." https://opendefinition.org/od/2.1/en/.

———. 2020. "Global Open Data Index." https://index.okfn.org/.

Haklay, Mordechai. 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, Polly, Adam Dennett, Thomas Russell, and Duncan Smith. 2018. "Colouring London–a Crowdsourcing Platform for Geospatial Data Related to London's Building Stock."

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage.

Langton, Samuel H, and Wouter Steenbeek. 2017. "Residential Burglary Target Selection: An Analysis at the Property-Level Using Google Street View." *Applied Geography* 86: 292–99.

Longley, Paul A. 2012. "Geodemographics and the Practices of Geographic Information Science." *International Journal of Geographical Information Science* 26 (12): 2227–37.

Loukaitou-Sideris, Anastasia. 1999. "Hot Spots of Bus Stop Crime: The Importance of Environmental Attributes." *Journal of the American Planning Association* 65 (4): 395–411.

Malleson, Nick, and Martin A Andresen. 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): 112–21.

Manovich, Lev. 2011. "Trending: The Promises and the Challenges of Big Social Data." *Debates in the Digital Humanities* 2: 460–75.

Mashhadi, Afra, Giovanni Quattrone, and Licia Capra. 2013. "Putting Ubiquitous Crowd-Sourcing into Context." In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 611–22.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

Newton, Andrew D, Shane D Johnson, and Kate J Bowers. 2004. "Crime on Bus Routes: An Evaluation of a Safer Travel Initiative." *Policing: An International Journal of Police Strategies & Management*.

Olmedilla, M, M Rocı́o Martı́nez-Torres, and SL Toral. 2016. "Harvesting Big Data in Social Science: A Methodological Approach for Collecting Online User-Generated Content." *Computer Standards & Interfaces* 46: 79–87.

Pridemore, William Alex, Matthew C Makel, and Jonathan A Plucker. 2018. “Replication in Criminology and the Social Sciences.” *Annual Review of Criminology* 1: 19–38.

Solymosi, Reka, and Kate Bowers. 2018. “The Role of Innovative Data Collection Methods in Advancing Criminological Understanding.” *The Oxford Handbook of Environmental Criminology* 210: 210–37.

Solymosi, Reka, Kate J Bowers, and Taku Fujiyama. 2018. “Crowdsourcing Subjective Perceptions of Neighbourhood Disorder: Interpreting Bias in Open Data.” *The British Journal of Criminology* 58 (4): 944–67.

Summers, Lucia, and Shane D Johnson. 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, Quan Hoang. 2017. “Open Data, Open Review and Open Dialogue in Making Social Sciences Plausible.” *Nature: Scientific Data Updates*. Available%20at%20SSRN:%20https://ssrn.com/abstract=3086667.