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

Supplementary materials

Samuel Langton (University of Leeds)

Reka Solymosi (University of Manchester)

# 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 et al., 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 et al., 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. Should you struggle to download any of the data during the exercise, or if the data sources have been updated since publishing, each one has been archived and is available on the corresponding GitHub page for this topic (<https://github.com/langtonhugh/osm_crim>).

## 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, and (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 datasets 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 to 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. Visiting this website, you will see a welcome message, and six tabs across the top 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 information that 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 take notes on what you think might be relevant for your analysis. For example, if we want to map crimes, we will 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!

To then 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 first using library().

library(readr)  
  
btp\_df <- read\_csv("data/2020-01-btp-street.csv")

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://github.com/langtonhugh/osm\_crim/raw/master/  
 data/2020-01-btp-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 kind 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 et al., 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’ and ‘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 offers 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.

[INSERT FIGURE 1 HERE]

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 will not 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 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 Jubilee Line. What do you think that URL will look like? Indeed, all you need to do is replace “bakerloo” with “jubilee”.

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 Jubilee line, one of the busiest on the network. Note how this is simply an amended version of the example provided by TfL in the API documentation.

library(jsonlite)  
  
api\_call <- fromJSON(readLines("https://api.tfl.gov.uk/line/jubilee/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 is what we might usually expect. 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 file 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 main 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 Open Street Map, 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 online: <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 need to specify a 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 Open Street Map 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? One way to do this is through a search term. Here, we want to select Greater London, so we can use the search term “greater london united kingdom” within the getbb() function. Using this function we can also specify what format we want the data to be in. In this case, we want a spatial object, specifically an sf polygon object (from the sf package), which we name bb\_sf.

library(osmdata)  
library(sf)  
  
bb\_sf <- getbb(place\_name = "greater london united kingdom",  
 format\_out = "sf\_polygon")

One downside of using search terms to define the bounding box, as with any search term, is that inconsistencies in terminology, wordings and spelling might lead to unexpected outputs (or none at all). Another way to obtain the bounding box is to manually specify the latitude and longitude coordinates. This requires a bit of pre-existing knowledge about your study region. A quick internet search, or exploration of the interactive Open Street Map platform (<https://www.openstreetmap.org/>), will give you an idea of the coordinates for your study region. For Greater London, we can specify our bounding box as follows, saving the coordinates to the object bb\_gl for later use.

bb\_gl <- c(-0.51037, 51.28676, 0.33401, 51.69187) # xmin, ymin, xmax, ymax

We can now move on to query 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(s) 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 are still working with spatial data, we stick to the sf format, for which the function is osmdata\_sf(). Here, we specify our bounding box as the manual coordinates contained in bb\_gl.

osm\_stat\_sf <- opq(bbox = bb\_gl) %>% # bounding box  
 add\_osm\_feature(key = 'public\_transport', value = 'station') %>% # select features  
 osmdata\_sf() # specify class

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

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 1736 at the time of writing. We also have around fifty 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 our study region mapped by Open Street Map 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 added information. Nevertheless, when the details are recorded, they provide rich insight and local knowledge that we may otherwise be unable to obtain.

One additional step needed is to select the stations which are relevant to us (i.e. they fall along the Jubilee Line). The variable line can help us with this. First, let’s explore this variable by using the table() function.

table(osm\_stat\_sf$line)

You will see that “Jubilee” appears several times under different line names. The main category (“Jubilee”) is the most common, but many other combinations exist (e.g. “Central;Jubilee”) for stations which sit on multiple lines. To get round this, and select all stations with “Jubilee” 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)  
  
osm\_jub\_sf <- osm\_stat\_sf %>%   
 filter(grepl("Jubilee", line))

This gives us Open Street Map data on all 27 stations on the Jubilee underground line. Now we have all our open datasets loaded into the R environment!

## Cleaning, wrangling, and visualizing 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 visualization. We’ve already loaded the osmdata and dplyr packages, but here, we will also make use of ggplot2, sf, 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(ggplot2)  
library(sf)  
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 Open Street Map 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\_jub\_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 (see Figure 2) using the ggplot2 package, which is compatible with sf objects.

ggplot(tfl\_jub\_sf) +  
 geom\_sf()

[INSERT FIGURE 2 HERE]

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 Jubilee underground line.

### Visualise to check and compare our datasets

We can now plot the Open Street Map station points, which are already spatial, over these station locations pulled from the TfL API, to see how the two datasets compare (see Figure 3).

ggplot() +  
 geom\_sf(data = tfl\_jub\_sf) +  
 geom\_sf(data = osm\_jub\_sf, col = "red", alpha = 0.5)

[INSERT FIGURE 3 HERE]

As we can see, the two data sources appear to be broadly comparable in terms of their spatial patterning. However, the exact location of stations sometimes differs between Open Street Map and TfL. Determining a single pinpoint location for underground stations is not straightforward, especially for larger stations which have multiple entrances, platforms, and walkway tunnels. For operational purposes, TfL might have a specific definition for the location of a station, such as its centroid, which has been systematically applied to create the coordinates. The data pulled from Open Street Map might have been compiled by a volunteer contributor with their own perception of how to define the location of a station, such as the ticket booth or the main entrance. This definition *might* be more theoretically meaningful than that used by TfL, since it could reflect a realistic ‘on the ground’ perception of the station’s built environment, and thus opportunities for crime. On the other hand, if the locations from Open Street Map were collated by multiple different contributors, each with their own idea of where each station lies on the line, we might conclude that the measure is unreliable. These kinds of questions demonstrate why critically engaging with open data sources is so crucial, but also challenging. We return to this at the end of the exercise.

Given the differences observed between the two open data sources, what might the impact be when studying crime in and around London Underground stations? 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 from the sf package that we used earlier, after first dropping crimes that have incomplete coordinates.

btp\_sf <- btp\_df %>%  
 drop\_na(Longitude, Latitude) %>%  
 st\_as\_sf(coords = c(x = "Longitude", y = "Latitude"), crs = 4326) %>%   
 st\_transform(27700)

We now have the recorded location of all crimes recorded by the British Transport Police during January, 2020. For the purposes of this demonstration, we can define crimes occurring as ‘in and around’ Jubilee 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 Tompson et al., 2015), but it does serve to facilitate this demonstration.

# Ensure each has the same CRS (BNG, 27700).  
osm\_jub\_sf <- st\_transform(osm\_jub\_sf, st\_crs(tfl\_jub\_sf))  
  
# Create buffers to define 'in and around'.  
osm\_buff\_sf <- st\_buffer(osm\_jub\_sf, dist = 50)  
tfl\_buff\_sf <- st\_buffer(tfl\_jub\_sf, dist = 50)  
  
# Count number of crimes recorded within each buffer.  
osm\_jub\_sf <- osm\_buff\_sf %>%   
 mutate(crimes = lengths(st\_intersects(osm\_buff\_sf, btp\_sf)))  
tfl\_jub\_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 buffers according to the crime count (see Figure 4). Note that we arrange the plots using syntax available using the patchwork package, and set the scales to be comparable using the lim argument within scale\_color\_viridis\_c().

p1 <- ggplot(data = osm\_jub\_sf) +  
 geom\_sf(mapping = aes(color = crimes), size = 2) +  
 labs(title = "Open Street Map") +  
 scale\_color\_viridis\_c(lim = c(0,42))  
   
p2 <- ggplot(data = tfl\_jub\_sf) +  
 geom\_sf(mapping = aes(color = crimes), size = 2) +  
 labs(title = "Transport for London") +  
 scale\_color\_viridis\_c(lim = c(0,42))  
  
p1 / p2

[INSERT FIGURE 4 HERE]

A number of factors emerge from this investigation. First, there is a disparity between the total number of crimes captured by the buffers in each dataset. Using the Open Street Map data, a total of 168 crimes were recorded across all stations on the line. Using the TfL data, only 136 were recorded. Thus, the volume of crime observed in and around Jubilee stations during January will differ depending on the open data source being used. This could have important implications for how community and policing resources are distributed between underground lines. Second, the spatial patterning of crime hotspots on the line differs considerably between the two data sources. Using the Open Street Map data, the final station on the line, Stratford, has the greatest concentration of crimes, whereas the TfL data suggests that the most problematic station is Green Park, in the city centre. Despite the minor differences in the locations of each station, the conclusions drawn are starkly different, and would determine the consistency of findings with existing research examining crime at underground stations (Ceccato et al., 2013).

We can explore Stratford station as a potential crime hotspot using some of the skills we have learnt so far. One useful first step might to visualize the built environment around the station, along with the crime buffers from Open Street Map and TfL respectively, and the crime locations themselves (see Figure 5). This way, we might find out why the TfL buffer is failing to capture any crimes, while the Open Street Map buffer is capturing an apparent crime hotspot. In doing so, it is worth remembering that crime locations in open police data are approximations of the real geography in order to ensure anonymity.

# Retrieve Stratford buffer from Open Street Map.  
stratford\_osm\_sf <- osm\_jub\_sf %>%   
 filter(name == "Stratford")  
  
# Do the same using TfL.  
stratford\_tfl\_sf <- tfl\_jub\_sf %>%   
 filter(commonName == "Stratford Underground Station")  
  
# Create a mini study area for our visualization, by getting the bounding box for the  
# Open Street Map (or TfL) buffer. Note that we transform to WGS 84 first in preperation  
# for the Open Street Map API.  
bb\_sf <- stratford\_osm\_sf %>%  
 st\_transform(crs = 4326) %>%   
 st\_bbox()   
  
# Retrieve the coordinates of the bounding box, and create a numeric list for use with  
# osmdata package.  
strat\_bb <- c(bb\_sf[[1]], bb\_sf[[2]], bb\_sf[[3]], bb\_sf[[4]])  
  
# Pull building footprints in this bounding box from the Open Street Map API.  
strat\_fp\_sf <- opq(strat\_bb) %>%  
 add\_osm\_feature(key = 'building') %>%   
 osmdata\_sf()  
  
# Get the polygons only.  
strat\_fp\_poly\_sf <- strat\_fp\_sf$osm\_polygons  
  
# Transform back to British National Grid to match our other data.  
strat\_fp\_poly\_sf <- st\_transform(strat\_fp\_poly\_sf, 27700)  
  
# Clip the crime data to the Open Street Map buffer. Remember that the TfL buffer  
# did not capture any crimes.  
strat\_osm\_btp\_sf <- st\_intersection(stratford\_osm\_sf, btp\_sf)  
  
# Visualize the building footprints, Open Street Map buffer, TfL buffer, and crime  
# locations. Note that the single point for the crimes is due to the approximation!  
# There are actually multiple crimes. Colors have been manually selected  
# from the viridis package to be easier to read by those with colorblindness.  
ggplot() +  
 geom\_sf(data = strat\_fp\_poly\_sf, fill = "black") +  
 geom\_sf(data = stratford\_osm\_sf, fill = "#440154FF", alpha = 0.3) +  
 geom\_sf(data = stratford\_tfl\_sf, fill = "#FDE725FF", alpha = 0.3) +  
 geom\_sf(data = strat\_osm\_btp\_sf, col = "#3CBC75FF", fill = "white") +  
 theme(axis.text = element\_text(size = 6))

[INSERT FIGURE 5 HERE]

This quick check has shed some light on what is happening. The buffer generated from the TfL station location (in yellow) covers a different part of the station to the Open Street Map buffer (in blue). The latter covers most of the main station building, while the former centers around the platforms to the north. Because the British Transport Police crime locations have been geocoded to a location in the main station building (in green) the TfL buffer has missed them. If we had used a larger buffer, it is likely that the TfL location would have captured the crime cluster, but then we would increase the risk of picking up crimes outside of the station. In many cases, there will not be an obvious ‘right or wrong’ answer to determine which open data source is most appropriate, but investigations like this one will help inform the decision. We might consider investigating other stations before deciding on which data to use, and ensure that findings are reported transparently with reference to these discussions.

Engaging with the creators of the data would also be an important step in answering the critical questions (‘Where’, ‘Why’, ‘Who’, ‘What’ and ‘When’) posed at the beginning of this chapter. For TfL, this might mean investigating the criteria used by their analysts to define the coordinates for stations. For Open Street Map, its online transparency means that we could find out which user(s) recorded the Jubilee station locations, and directly contact this community of volunteers to find out how station locations were recorded and why.

# Software

As noted in the tutorial, a number of R packages were used to create the material in this chapter. Packages *readr* (Wickham et al., 2018) and *jsonlite* (Ooms, 2014) were used for reading and parsing data, *osmdata* for querying the Open Street Map API (Padgham et al., 2017), *sf* for handling and visualizing spatial data (Pebesma, 2018), *dplyr* and *tidyr* for wrangling data (Wickham et al., 2020; Wickham, 2020), *ggplot2* for visualization (Wickham, 2016) and *patchwork* for arranging plots (Pedersen, 2020). The draft versions of the book chapter and supplementary material were prepared using *rmarkdown* (Xie et al., 2018).

# Learn more about it

A useful resource going forward is the *Open Data Handbook* which can be found online for free (<https://opendatahandbook.org/>). It provides introductory information (e.g. definitions), advice (e.g. legal, technical details), case studies which showcase the societal value of open data, and an e-library of papers, presentations and videos on open data. It is one of many projects organized by the *Open Knowledge Foundation*. Another comprehensive and free online resource is the *Open Data Institute* website (<https://theodi.org/>). Their mission is to demonstrate the positive value of open data and advocate for its ethical creation and usage by working with companies and governments around the globe. For resources geared towards R, *rOpenSci* (<https://ropensci.org/>) develop packages and offer an infrastructure and community to promote the accessible usage of R for open scientific research. The official *RStudio* webpage (<https://rstudio.com/about/>) also provides a wealth of information and freely available resources for making the most of the software.

# References

Ceccato, V., Uittenbogaard, A., & Bamzar, R. (2013). Security in stockholm’s underground stations: The importance of environmental attributes and context. *Security Journal*, *26*(1), 33–59.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2016). *Big data and social science: A practical guide to methods and tools*. crc Press.

Loukaitou-Sideris, A. (1999). Hot spots of bus stop crime: The importance of environmental attributes. *Journal of the American Planning Association*, *65*(4), 395–411.

Newton, A. D., Johnson, S. D., & Bowers, K. J. (2004). Crime on bus routes: An evaluation of a safer travel initiative. *Policing: An International Journal of Police Strategies & Management*.

Olmedilla, M., Martı́nez-Torres, M. R., & Toral, S. (2016). Harvesting big data in social science: A methodological approach for collecting online user-generated content. *Computer Standards & Interfaces*, *46*, 79–87.

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO]*. <https://arxiv.org/abs/1403.2805>

Padgham, M., Rudis, B., Lovelace, R., & Salmon, M. (2017). Osmdata. *The Journal of Open Source Software*, *2*(14). <https://doi.org/10.21105/joss.00305>

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. <https://doi.org/10.32614/RJ-2018-009>

Pedersen, T. L. (2020). *Patchwork: The composer of plots*. <https://CRAN.R-project.org/package=patchwork>

Tompson, L., Johnson, S., Ashby, M., Perkins, C., & Edwards, P. (2015). UK open source crime data: Accuracy and possibilities for research. *Cartography and Geographic Information Science*, *42*(2), 97–111.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>

Wickham, H. (2020). *Tidyr: Tidy messy data*. <https://CRAN.R-project.org/package=tidyr>

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. <https://CRAN.R-project.org/package=dplyr>

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*. <https://CRAN.R-project.org/package=readr>

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>