Crowdsourcing public perceptions of space and safety

David Buil-Gil and Reka Solymosi
Department of Criminology, University of Manchester, UK

24/04/2020

Abstract

Insert abstract here
Keywords: Fear of crime, perceived safety, crime mapping, open data, GIS, Atlanta

Full reference: Buil-Gil, D., & Solymosi, R. (2020). Crowdsourcing public perceptions of space and safety. In E. Groff & C. Haberman (Eds.), *The Study of crime and place: A methods handbook*. Temple University Press.

Contact details: David Buil-Gil. G18 Humanities Bridgeford Street Building, Cathie Marsh Institute for Social Research, University of Manchester. E-mail address: *david.builgil@manchester.ac.uk*

ORCID IDs: David Buil-Gil: 0000-0002-7549-6317. Reka Solymosi: 0000-0001-8689-1526.

1. Introduction

Crowdsourcing refers to the practise of enlisting the knowledge, experience or skills of a large number of citizens (*the 'crowd'*) to achieve a common goal or cumulative result, usually via a platform powered by digital technologies, mobile phones, social media or a website (Howe 2006). Digital platforms allow recording large volumes of data in relatively little time at a very small cost, which explains why data generated through crowdsourcing is currently utilized for a variety of function ranging from academic research to policy making and emergency management (Brabham 2008; Goodchild 2007). As an example, during the 2007-2009 wildfires in the Santa Barbara area, California, residents shared their real-time knowledge about the location of fires and emergency shelters via various online forums and websites, which proved to be an invaluable source of information for disaster response (Goodchild and Glennon 2010). Similarly, crowdsourcing projects have been deployed to harness people's experiences with crime and their perceptions about space and safety (e.g., Solymosi and Bowers 2018; Williams, Burnap, and Sloan 2017). In this chapter we review some published literature about the use of crowdsourcing for criminological research, discuss the main strenghts and limitations of data produced from crowdsourcing platforms, and present an step-by-step exemplar study in `R` software (R Core Team 2020) using crowdsourced perceptions of safety in Atlanta, Georgia.

In criminological research, crowdsourcing analysis has been primarily used to harness data about various forms of crime and antisocial behavior and to process information about citizens' perceptions and emotions about crime, thus allowing researchers to devise new explanations of crime and perceived safety. On one hand, open data recorded from social media and online forums enables detecting various forms of online crimes (e.g., hate speech towards minority groups; Miró-Llinares, Moneva, and Esteve 2018), and even associate those forms of online hate speech with offline racially and religiously aggravated crimes (Williams et al. 2020). Moreover, Williams, Burnap, and Sloan (2017, 334) argue that "open-source communications, in particular from Twitter, have potential for measuring the breakdown of social and physical order at the borough level", which is also shown by Erete et al. (2016). Thus, some criminologists use large crowdsourced datasets to detect and explain criminal activity.

On the other hands, those researchers interested in the public perceptions about space and safety explore the use of volunteered crowdsourced information to explain the citizens' perceptions and emotions about crime. Criminologists have long known that public emotions about crime are not always explained by the prevalence and harms of crime, but instead fear of crime emotions are the result of individual predispositions to experience negative emotions about crime, which become fear of crime episodes under the presence of certain social and situational influences (Gabriel and Greve 2003; Hale 1996). The public perceptions and emotions about crime have traditionally been analysed by using surveys and interview-type qualitative approaches (see Gabriel and Greve 2003; Warr 2000), but these methods are costly and may be limited to capture the time and context-specific emotional reactions of fear of crime as well as the citizens' behavioural responses to such emotions of fear (e.g., avoidance behaviors, acquiring alarm systems or weapons). As an alternartive, some researchers have endorsed the use crowdsourcing and app-based tool to record data about the places and times in which episodes of fear of crime are more frequent, in order to fully conceptualize and operationalize the fear of crime as a function of individuals and their immediate environment (Solymosi and Bowers 2018; Solymosi et al. 2020).

To mention only a few examples of crowdsourcing platforms deployed to record data about perceptions and emotions about crime, Hamilton et al. (2011) developed a mobile phone app to record public perceptions of crime on public transportation in Melbourne, Australia. Similarly, Solymosi, Bowers, and Fujiyama (2015) designed an app and asked participants to report their worry about crime, which allows authors to map the users' fear pf crime across the different areas of London, UK. Salesses, Schechtner, and Hidalgo (2013) designed and recorded data from the Place Pluse platform, which asks participants to choose 'which place looks safer' between two images taken from Google Street View. Then Salesses, Schechtner, and Hidalgo (2013) produced a map of perceived safety in New York, USA. Birenboim (2016) developed a mobile app to record data about the perceptions of security of attendees at a music festival in Jerusalem, Israel. Gómez et al. (2016) designed a collaborative web-based tool that allowed the citizens of Bogota, Colombia, to report those areas in which they feel less safe. And Solymosi, Bowers, and Fujiyama (2017) analysed secondary data recorded by FixMyStreet, an online problem reporting website, to examine perceptions of neighborhood disorder in London, UK. These are only a few examples, but there are many others crowdsourcing platforms that have been designed and utilized to study the fear of crime (see Solymosi et al. 2020).

2. Strengths and weaknesses of crowdsourced data

Although crowdsourced data about public perceptions of space and safety have some key strengths over data recorded by traditional survey methods (e.g., precise spatial data, information about immediate environmental variables, reduced cost), their unique mode of production is also associated to certain limitations or weaknesses that, if uncontrolled, may affect the validity of such measures and the reliability of final results (Buil-Gil, Solymosi, and Moretti 2020; Elliott and Valliant 2017). Amongst those limitations identified by researchers using crowdsourced data to study the fear of crime, the ones that are most commonly mentioned are related to the participation inequality arising from self-selection in crowdsourced samples, the difficulty to interpret results, and the under-representation of certain areas and times (Solymosi et al. 2020). Others also identify that the number of participants in crowdsourcing projects tends to decrease over time (i.e., participation decrease), and some platforms have difficulties to engage participant and can only record small samples (e.g., Blom et al. 2010). We will briefly review some of these strengths and weaknesses of crowdsourced data and illustrate some of them with data about perceptions of safety in Atlanta.

Solymosi et al. (2020) STRENGTHS

Capture the spatial-temporal specific nature of attitudes and emotions towards crime. 23/27 Record data on individual variables and specific types of fear/disorder. 12/27 Record data on architectural features. 20/27 Reduced cost of data collection. 11/27

WEAKNESSES

Self-selection bias: males tend to be more represented than females, and young citizens are overrepresented compared to older groups Chataway et al. (2017) check Salesses, Schechtner, and Hidalgo (2013)

Participation inequality: FixMyStreet data, one fourth of all reports had been produced by one percent of users, while 73% of people in the sample contributed only once (Solymosi, Bowers, and Fujiyama (2017)).

Under-representation of certain areas

Participation decrease, sample attrition over time Solymosi et al. (2020) Blom et al. (2010)

Repeatedly asking about fear might increase/cause fear

Lack of temporal variability in some web-based measures

Validity of measures

3. Crowdsourcing perceptions of safety: Step-by-step example in R

In order to illustrate the use of crowdsourced data in criminological research, we present below an exemplar study using data recorded by the Place Pulse 2.0 platform (Salesses, Schechtner, and Hidalgo 2013). This section will introduce the Place Pulse project and provide R codes to download, explore and clean this source of crowdsourced data. Then, we will analyse the spatial distribution of crowdsourced perceptions of space and safety in Atlanta, Georgia, and illustrate with examples some known issues of crowdsourced data (i.e., participantion inequality, participation decrease and measure validity).

3.1 The Place Pulse project

Place Pulse 2.0 was an online crowdsourcing platform designed to record data about citizens' perceptions of safety, beauty, wealth, liveability, boredom and depression. Two images obtained were shown to participants, who then were asked 'Which place looks safer?' (see Figure 1). The platform could also be used to report which of the two images looked wealthier, more beautiful, more boring, livelier or more depressing, but we will focus on the 'safer' answers in this chapter. Images were selected randomly from Google Street View across 56 cities from 28 countries, and these captions were originally taken between 2007 and 2012. The platform recorded all responses in a public dataset, but responders did not provide any further information about themselves, which means that we do not know their social and demographic characteristics. Place Pulse used to be hosted in an open website (http://pulse.media.mit.edu/) and anyone could use it, but the platform was closed in 2019. We have been granted access to all the data recorded between the 28th May 2013 and the 22nd August 2019 to write this chapter, which has been uploaded onto an open repository with consent of the data producers (Salesses, Schechtner, and Hidalgo 2013).

3.2 Download and explore Place Pulse data

We have saved all Place Pulse data in a data repository on FigShare. You can download this directly into R by using the `read.csv()` function. It is a large file so it may take some minutes to read in.

```
pp_data <- read.csv('https://ndownloader.figshare.com/files/21739137')
```

This data set includes 17 variables, but we will only use some of them. A unique identification code was given to each participant (*'voter_uniqueid'*) and image (*'place_id_left'* for images in the left side of the pairwise comparison and *'place_id_right'* for images in the right part). The columns *'place_name_left'* and *'place_name_right'* specify the city of each photograph. The column *'choice'* shows if the user perceived the image in the left or right to be 'safer' (participants could also answer 'equal'), and the column *'study_question'* allows studying perceptions about different variables (e.g., safety, wealth, beauty). The columns *'day'* and *'time'* specify the moment when each vote took place, and the columns *'long_right'*, *'lat_right'*, *'long_left'* and *'lat_left'* show the longitude and latitude of both photographs.

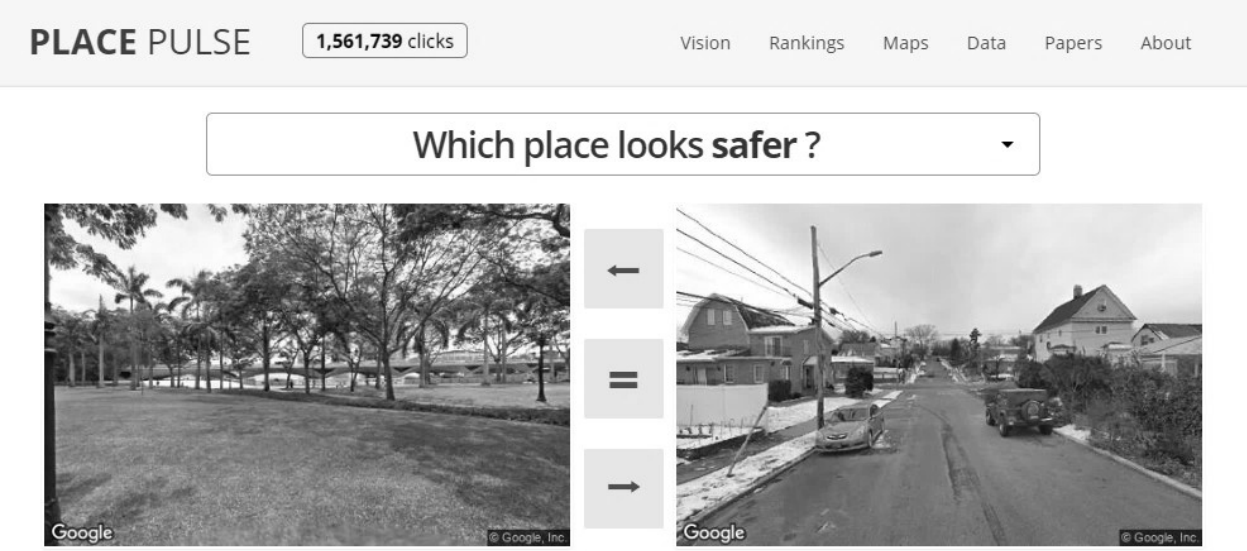


Figure 1: Figure 1: Place Pulse website

We can begin by examining which cities were more frequenty assessed within the Place Pulse platform. This can be checked for images in the left side of the pairwise comparison first, and then for images in the right. We can use the `group_by()` and `summarize()` functions from `dplyr` package (Wickham et al. 2020) to check this:

```
pp_data %>%
  group_by(place_name_left) %>% # categories based on cities on the left
  summarize(Count = n()) %>% # count number of units in each category
  top_n(3) # print 3 most frequent categories
```

```
## # A tibble: 3 x 2
##   place_name_left Count
##   <fct>           <int>
## 1 Atlanta         56992
## 2 Berlin          55265
## 3 Tokyo           53817
```

Atlanta was the city with a largest number of votes amongst those images that appeared in the left part of the comparison, but we can also check this for those images shown in the right side:

```
pp_data %>%
  group_by(place_name_right) %>% # categories based on cities on the right
  summarize(Count = n()) %>% # count number of units in each category
  top_n(3) # print 3 most frequent categories
```

```
## # A tibble: 3 x 2
##   place_name_right Count
##   <fct>            <int>
## 1 Atlanta          57140
## 2 Berlin           55583
## 3 Tokyo            53514
```

Atlanta was indeed the city with the largest number of votes. We can also check which variables (e.g., safety, beauty, wealth) were more frequently assessed by participants:

```
pp_data %>%
  group_by(study_question) %>% # categories based on study questions
  summarize(Count = n()) %>% # count number of units in each category
  arrange(desc(Count)) # reorder in descending order
```

```
## # A tibble: 7 x 2
##   study_question   Count
##   <fct>            <int>
## 1 safer           509961
## 2 livelier        366802
## 3 more beautiful  220604
## 4 wealthier       174758
## 5 more depressing 149355
## 6 more boring     144060
## 7 <NA>               183
```

We see that safety was the most commonly assessed variable, with 509,961 votes in total. In this chapter we will examine reports of safety in the city of Atlanta. Before analysing the data, however, we can also examine if participants were more inclined to vote for images in the left or right part of the platform. In other words, we analyse if responses were biased by the position in which images were shown on the website.

```
pp_data %>%
  group_by(choice) %>% # categories based on vote (right, left or equal)
  summarize(Count = n()) %>% # count number of units in each category
  top_n(3) # print 3 most frequent categories
```

```
## # A tibble: 3 x 2
##   choice  Count
##   <fct>   <int>
## 1 equal  206147
## 2 left   668680
## 3 right  690792
```

The frequency of votes for left and right options is very similar, and thus we can conclude that the position of the image on the web does not appear to have an affect on participants' votes.

3.3 Cleaning Place Pulse data

When it comes to crowdsourced data, you will have to be an expert data wrangler to make sure you can get the data to behave like you want it to - in other words, to make the data available in a format that allows you to answer your research questions. For example, in this case, we want to map the perceived safety of areas in Atlanta. To do this, first we have to select the area of Atlanta and the votes about safety.

Atlanta is the capital city of the State of Georgia, in the US. In 2018, its estimated papulation was close to 500,000 residents, and thus it is the 37th most populated city in the United States. There are two main reasons why we are conducting this exemplar study in Atlanta: first, as shown above, it was the city with the largest number of votes in the Place Pulse platform; and second, there is available open data about the social, demographic and crime characteristics of Atlanta neighborhoods, which can be easily accessed to explain the patterns observed in Place Pulse data. Moreover, various papers have analysed the predictors of

crime and fear of crime in this city, which can be used to interpret our results (see McNulty and Holloway 2000; Tester et al. 2011)

We can use the function `filter()` from `dplyr` to create a new dataframe that includes those pairwise comparisons in which either the image of the right or the image of the left, or both, are from Atlanta.

```
# select cases in which the image of the right or left is from Atlanta
pp_atl <- pp_data %>%
  filter(place_name_right == "Atlanta" | place_name_left == "Atlanta")
```

We will also focus on the ratings of areas as safer, as we are interested in people's perceptions of place and safety:

```
pp_atl_s <- pp_atl %>%
  filter(study_question == "safer") # select votes of 'safer'
```

You can see now that we have a dataframe of 37214 votes about the safety of places in Atlanta.

We are interested in analysing the proportion of 'safer' votes in each neighbourhood of Atlanta. In order to assign each photograph to their neighbourhood, we need to create two new columns that specify the longitude and latitude of the image being assessed. We will also create a new column that details whether each participant voted that the image of Atlanta was 'safer' or 'not safer' (i.e., less safe or equal) than the other picture. Some pairwise comparisons, however, assessed two different images from Atlanta, which means that we will need to duplicate these votes to account for both the image on the right and left of the pariwise comparison. First, we want to know the number of comparisons in which both images are from Atlanta. We can use the function `filter()` from `dplyr`, which we have already used above.

```
# create dataset of votes in which both images are from Atlanta
pp_atl_s_dup  <- pp_atl_s %>%
    filter(place_name_right == "Atlanta" & place_name_left == "Atlanta")

# print the count of votes as a result
pp_atl_s_dup %>%
    summarize(count = n())
```

```
##   count
## 1   678
```

In total, 678 pairwise comparisons are based on two images from Atlanta, whereas 36,536 compare one image from Atlanta with a picture from any other city. We will duplicate those comparisons in which both images were taken in Atlanta and attach them to two new datasets (one to assess the images on the right, and the other to report the images on the left). For now, we delete duplicated cases from the main dataframe by using the `anti_join()` function from `dplyr`, which searches those duplicated units that exist in the newly created dataset `pp_atl_dup` and deletes them from the main dataset of votes. We will merge all units together once all the data has been cleaned.

```
# duplicate the new dataset
pp_atl_s_dup2 <- pp_atl_s_dup

# delete duplicated votes from main dataset
pp_atl_s <- pp_atl_s %>%
  anti_join(x = pp_atl_s, y = pp_atl_s_dup, by = "X")
```

Now, the main dataset includes only those pairwise comparison in which only one image is from Atlanta. We create two new columns that specify the coordinates of the photograph from Atlanta. We use the `mutate()` function from `dplyr` to create the new columns and the `if_else()` function to copy the coordinates of the image of the left or right depending on whether the left photoraph was from Atlanta or not.

```
# copy coordinates from left image if it is from Atlanta, otherwise copy from right image
pp_atl_s <- pp_atl_s %>%
    mutate(long_Atl = if_else(place_name_left == "Atlanta", long_left, long_right),
           lat_Atl  = if_else(place_name_left == "Atlanta", lat_left , lat_right))
```

Remember that we had previously created two new datasets with those votes in which both images are from Atlanta. The first dataset (`pp_atl_s_dup`) will be used to assess the images in the left, while the second dataset (`pp_atl_s_dup2`) refers to the images in the right. We can then allocate the coordinates of the left image to first dataset of votes in which both images are from Atlanta:

```
# copy coordinates from left image to new columns
pp_atl_s_dup <- pp_atl_s_dup %>%
  mutate(long_Atl = long_left,
         lat_Atl  = lat_left)
```

And then we allocate the coordinates of the right image to the second dataset of pairwise comparisons between Atlanta pictures:

```
# copy coordinates from right image to new columns
pp_atl_s_dup2 <- pp_atl_s_dup2 %>%
  mutate(long_Atl = long_right,
         lat_Atl  = lat_right)
```

Before merging all the data into a single dataset, we will also create a new column that distinguishes those images of Atlanta that were assessed as 'safer' from those reported as 'less safe' or 'equal'. We do this first in the main dataset (which only includes pairwise comparisons with one picture from Atlanta), by checking if the choice of each vote (i.e., 'left', 'right', or 'equal') corresponds to the position of the image from Atlanta. We assign a 1 if the user choose the image of Atlanta as 'safer', whereas a 0 is assigned when the image of a different city was chosen to be 'safer' and when users voted 'equal'. We can also use the `mutate()` and `if_else()` functions from `dplyr` here.

```
# if left image is from Atlanta and user voted for left image, assing 1, otherwise 0
# if right image is from Atlanta and user voted for right image, assign 1, otherwise 0
pp_atl_s <- pp_atl_s %>%
  mutate(win = if_else((place_name_left  == "Atlanta" & choice == "left") |
                       (place_name_right == "Atlanta" & choice == "right"), 1, 0))
```

Similarly, in the `pp_atl_s_dup` dataset, we assign a 1 to those cases in which participants voted for the left image as 'safer' and a 0 otherwise, given that this dataset had been previously created to assess left images in cases when both images were from Atlanta. And we do the same with the second dataset (`pp_atl_s_dup2`) that compares two images from Atlanta, which in this case refers to the image in the right.

```
# if participant voted for left image, assign 1, otherwise 0
pp_atl_s_dup <- pp_atl_s_dup %>%
  mutate(win = if_else(choice == "left", 1, 0))

# if participant voted for righ image, assign 1, otherwise 0
pp_atl_s_dup2 <- pp_atl_s_dup %>%
  mutate(win = if_else(choice == "right", 1, 0))
```

Now that our dataset has been cleaned and is ready to be analysed, we can merge all the data together wit the `rbind()` function.

```
pp_atl_s <- rbind(pp_atl_s, pp_atl_s_dup, pp_atl_s_dup2)
```

We have a dataframe of 37892 votes about the safety in Atlanta that is ready to be analysed.

3.4 Map Place Pulse data

It is possible to use mapping techniques learned in other chapters (LINK WITH MAPPING CHAPTER) to map the crowdsourced data (see also Solymosi, Bowers, and Fujiyama 2015). We will be using the `sf` and `ggplot2` libraries in order to create a map of perceived safety of built environment across the areas of Atlanta.

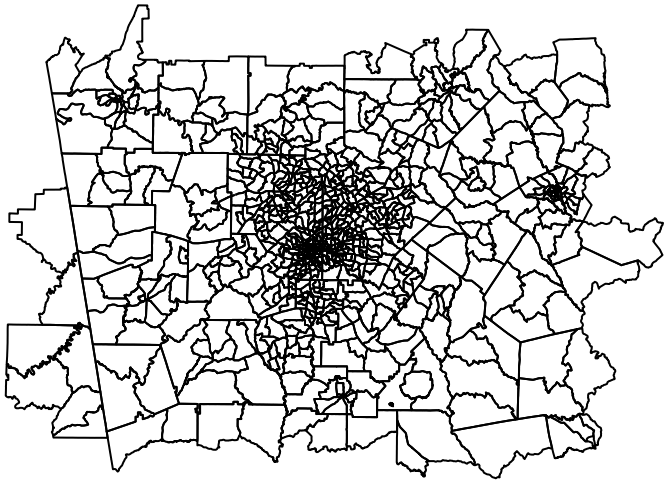
First, acquire a shapefile for Atlanta. Let's use the census tracts shapefile from Harvard University. You can go on their website to find out more about this boundary data: https://worldmap.harvard.edu/data/geonode:Atlanta_Census_Tracts_SHL. We can download the shapefile directly using their Application Programme Interfact (or API) - this is something discussed in greater detail on the chapter on Open Data (CHAPTER REF LANGTON AND SOLYMOSI, 2020). For now, you can just use the code below, with the `st_read()` function in the `sf` package (Pebesma 2020):

```
atl <- st_read("http://worldmap.harvard.edu/download/wfs/1824/json?outputFormat=json&service=WFS&reques
```

```
## Reading layer `OGRGeoJSON' from data source `http://worldmap.harvard.edu/download/wfs/1824/json?outpu
## Simple feature collection with 799 features and 29 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -85.65203 ymin: 32.96982 xmax: -82.94681 ymax: 34.58748
## CRS:            4326
```

We can see what this file looks like by using the `plot()` function to plot the geometry of the 'atl' object we created, called with the `st_geometry()` function:

```
plot(st_geometry(atl))
```



Now, to be able to plot the safety votes on this map, we first need to make our votes a spatial object, by specfying that the "long_Atl" and "lat_Atl" columns contain our longitude and latitude information. We use the `st_as_sf()` function for this:

```
points_atl_s <- st_as_sf(pp_atl_s, coords = c("lat_Atl", "long_Atl")) #geocode votes
```

In order to plot both these spatial layers (i.e., votes recorded from Place Pulse and Atlanta census tracts) on the same map, their coordinate reference systems (CRS) need to match. We can check these with the `st_crs()` function:

```
st_crs(points_atl_s) == st_crs(atl) #check if CRS is the same of both layers
```

```
## [1] FALSE
```

We function tells us that it is 'false' that both CRS are not equal, but we can change this with the following line of code:

```
st_crs(points_atl_s) <- st_crs(atl)
```

Now, if we check, they should have the same CRS:

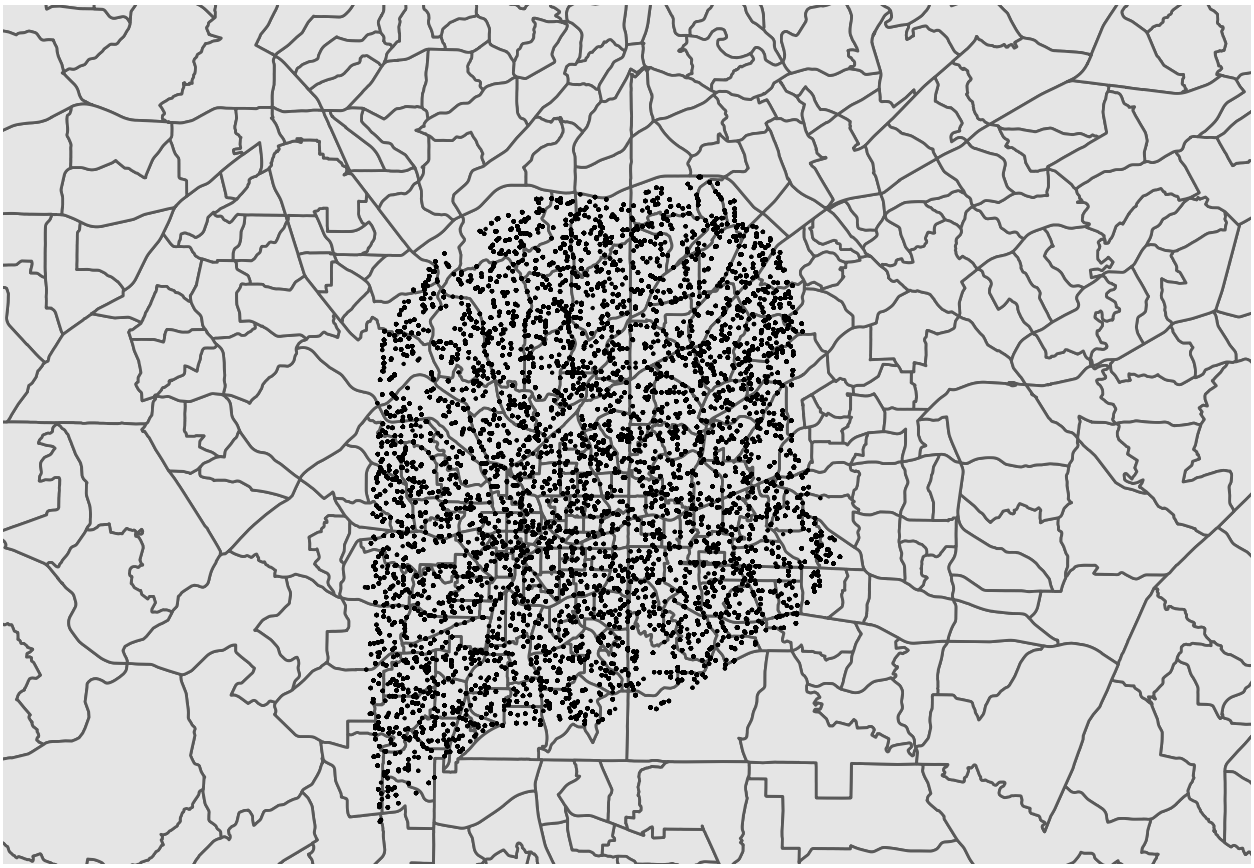
```
st_crs(points_atl_s) == st_crs(atl) #check if CRS is the same of both layers
```

```
## [1] TRUE
```

Thus, we can now map our data. See (LINK WITH MAPPING CHAPTER) for more information about crime mapping.

```
map <- ggplot(data = atl) + geom_sf() + theme_void() +
       coord_sf(xlim = c(-84.7, -84), ylim = c(33.6, 34),
                expand = FALSE) #create map

map + geom_point(data = pp_atl_s, aes(x = lat_Atl, y = long_Atl),
                 size = .1) #plot map with points
```



This is a very busy map. Maybe instead we want to get some sort of average score for each census tract. We then calculate the proportion of 'safer' responses in each area (see Buil-Gil, Solymosi, and Moretti 2020). Note that Salesses, Schechtner, and Hidalgo (2013) suggest computing a Q-score per image corrected by the "win" and "loss" ratio of all photographs with which it is compared, but for the purpose of this chapter we will compute a simple proportion that will allow us to directly analyse the geographical distribution of perceived safety.

```
points_atl_s_nhood <- st_intersection(atl, points_atl_s) %>%
                      group_by(TRACT) %>%
                      summarise(winscore = mean(win, na.rm = TRUE),
                                num_votes = n())
```

And we can add the average score of 'safer' reponses per area to the original shapefile of census tracts. We will also delete those census tracts in which we do not have any vote of perceived safety.

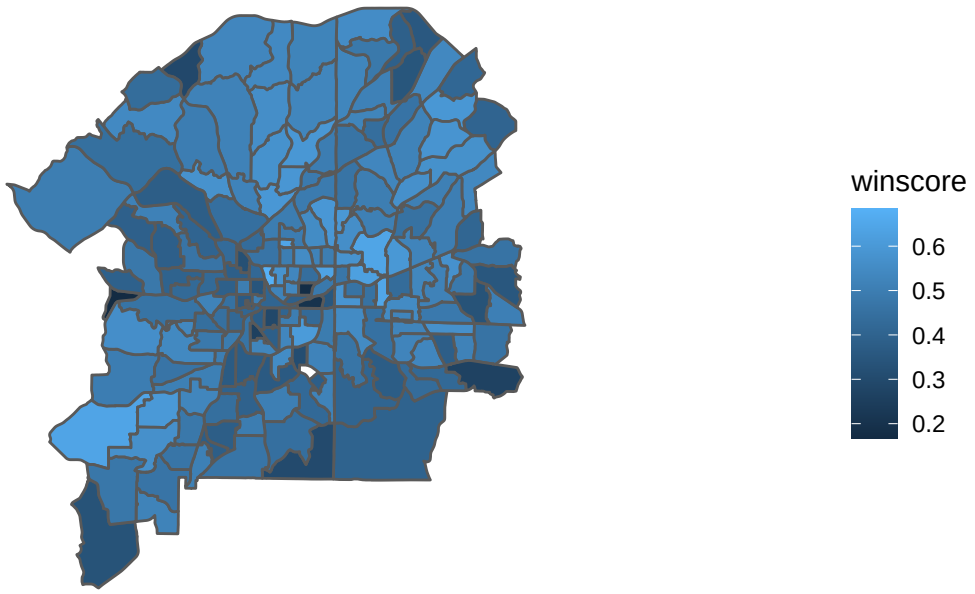
```r
st_geometry(points_atl_s_nhood) <- NULL

# merge census tracts and Place Pulse votes based on 'TRACT' column
atl_pp_wins <- left_join(atl, points_atl_s_nhood, by = c("TRACT" = "TRACT"))

# delete census tracts with 0 votes (NAs)
atl_pp_wins <- atl_pp_wins[!is.na(atl_pp_wins$winscore), ]
```

Finally, we can plot the propotion of 'safer' votes in each census tract.

```r
ggplot(data = atl_pp_wins) +
  geom_sf(aes(fill = winscore)) +
  coord_sf(xlim = c(-84.7, -84), ylim = c(33.5, 34), expand = FALSE) +
  theme_void()
```



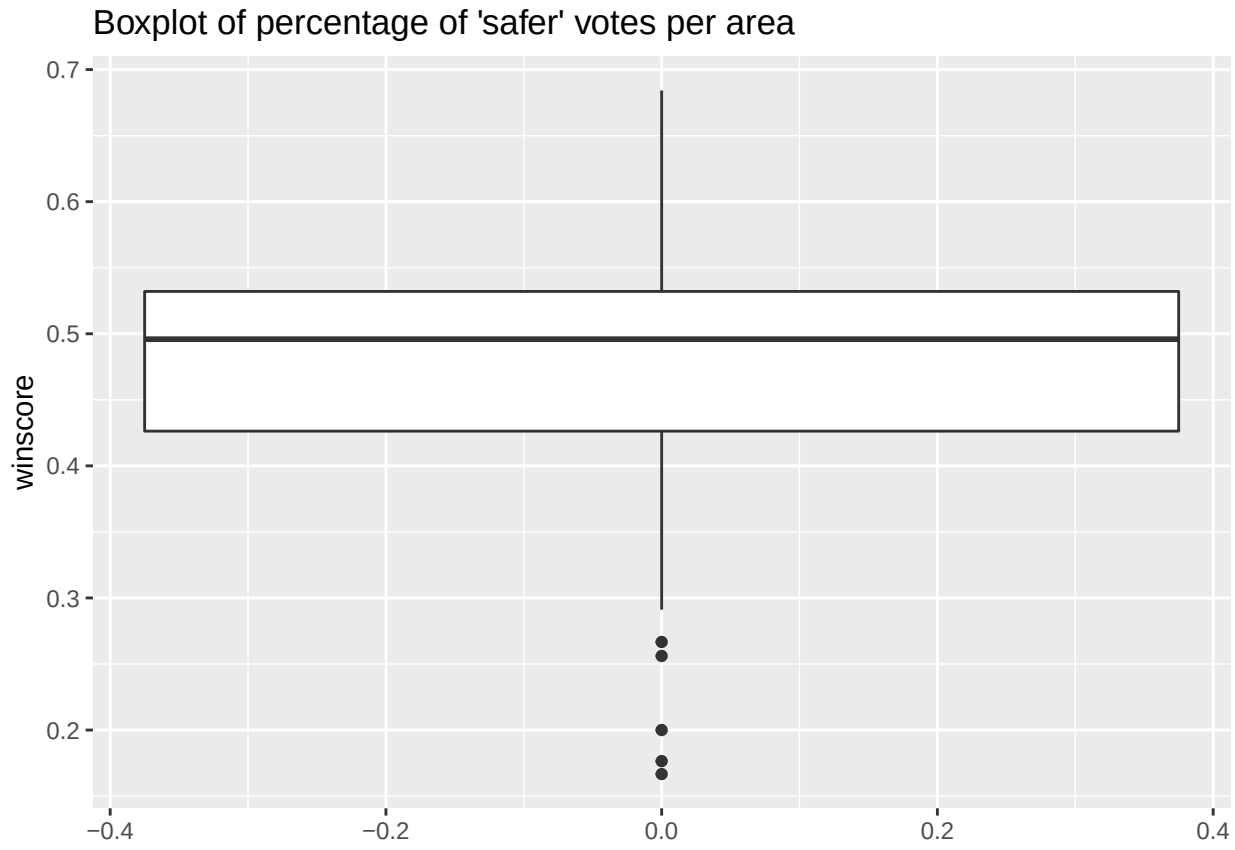
There are many more things one can do with these data. For example, we could look at the descriptive statistics and boxplot of the proportion of 'safer' votes in each area:

```r
summary(atl_pp_wins$winscore) # descriptive statistics: proportion 'safer' votes per tract
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1667  0.4263  0.4959  0.4798  0.5321  0.6842
```

```r
p <- ggplot(atl_pp_wins, aes(y = winscore))

p + geom_boxplot() + ggtitle("Boxplot of percentage of 'safer' votes per area") # boxplot
```



You can do much more, but here we will focus on the specific issues to explore due to the crowdsourced nature of these data.

3.5 Exploring known issues of crowdsourced data within Place Pulse

In section 2, we mentioned a few issues that are usually present in crowdsourced data, and which are important to keep in mind when ussing these data in criminological research. Here we explore whether some of these issues are present in data recorded from Place Pulse, and what that might mean for any conclusions we draw from our analyses. We shall keep in mind that the Place Pulse project did not record information about participants demographic characteristics, and thus we cannot direclty examine the self-selection biases that may affect this dataset (Buil-Gil, Solymosi, and Moretti 2020). Nevertheless, the sample's self-selection biases should be checked when possible. For instance, the first edition of Place Pulse, which was designed to register some demographic variables from participants, identified that 76.0% of those who reported their gender were males, and only 21.1% were females, and the median self-reported age was 28 years (Salesses, Schechtner, and Hidalgo 2013). Here we examine data from Place Pulse 2.0, which does not record social and demographic variables from participants, but we will examine whether our sample size is affected by other issues such as participation inequality, under-representation of certain areas and participation decrease. We will also study the validity of these crowdsourced data as a measure of perceived safety.

3.5.1 Participation inequality (supercontributors)

Crowdsourced data is usually affected by a few number of supercontributors that produce most votes. In order to check if our dataset is affected by this, we first need to create a new dataframe showing the number of votes that each study participant had made. To do this, we will use code from the `dplyr` library:

```
voter <- pp_data %>%
  group_by(voter_uniqueid) %>%
  summarise(num_votes = n())
```

If you want, you can have a look at this new dataframe using the `View()` function. If you do this, you might see, we have some very active participants. The top voter, for instance, has made 7168 votes on places. That is some very prolific participation! On the other hand, you can also see that 7494 of the participants made only one vote. We are definitely seeing signs of participation inequality in these data.

In fact, we can examine how much of the votes are produced by these "supercontributors". For example, we can assess the proportion of votes made by the top 1% of voters. We can do this using the `subset()` and `quantile()` functions:

```
# subset top 1% of most prolific participants
top_1percent <- subset(voter, num_votes > quantile(num_votes, prob = 1 - 1/100))
```

We can see that this new dataframe contains 954 people, who are our top 1% contributors to the Place Plus dataset. We will now examine how much of the total number of votes are generated by the top 1% of users:

```
sum(top_1percent$num_votes) / sum(voter$num_votes) * 100
```

```
## [1] 17.87468
```

That is a lot: 17.87% of the votes are made by the top 1% of contributors.

Activity 1 Can you tell me what proportion of the votes were made by the top 10% of participants? What about the top 25%?

3.5.2 Quantifying participation inequality

One way to quantify the extent to which participation inequality exists in our data is by using a Gini index, and visualising it using a Lorenz curve. The Gini index (or Gini ratio) is a measure of statistical dispersion intended to measure inequality (Gastwirth 1972). Although it tends to be used to examine income inequality, it has also been frequently used to assess participation inequality in crowdsourcing platforms (see Solymosi, Bowers, and Fujiyama 2017; Solymosi and Bowers 2018). Similarly, the Lorenz curve is a visual representation of inequality. For this you will need the library `ineq` (Zeileis and Kleiber 2015) so if you do not already have this you must install with the command:

```
install.packages("ineq")
```

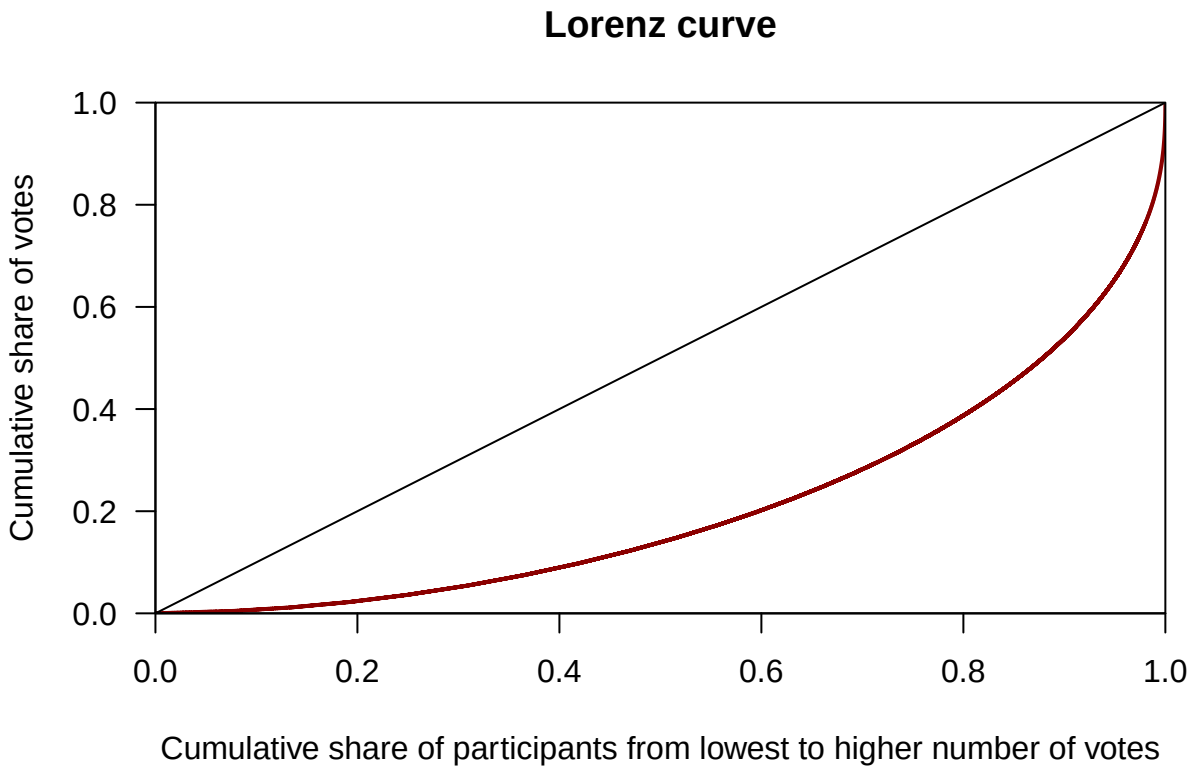
Then you can load this library and calculate the index using the `Gini()` function:

```
Gini(voter$num_votes)
```

```
## [1] 0.5777568
```

Remember that a score of 0 is perfect equality (everyone makes equal number of votes), while 1 is perfect inequality (only one person making all the reports, and no one else making any). Our answer of 0.58 shows some serious inequality. To put this into context, in 2017, according to the OECD, income inequality in the United States of America showed a Gini coefficient of 0.39. To visualize this we can use a Lorenz curve using the `plot()` and `Lc()` functions:

```
plot(Lc(voter$num_votes),
     xlab = "Cumulative share of participants from lowest to higher number of votes",
     ylab = "Cumulative share of votes", col = "darkred", lwd = 2)
```



The Lorenz curve (red line) above shows how the top few percent of reporters contribute the majority of the reports. If we had perfect equality, we would expect to see the red line align perfectly with the black line with the slope of 1. With this information we can now quantify how severe the participation inequality is in our data, and compare with other crowdsourced data for context and understanding.

3.5.3 Under-representation of certain areas

It is also important to consider variation in the sample size of number of votes in each census tract. We can analyse it certain areas are under-represented in our dataset by using the `summary()` function.

```
summary(atl_pp_wins$num_votes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   79.25  137.50  183.33  241.25 1053.00
```

Whereas the average sample size per area is quite large,

CONTINUE HERE!!

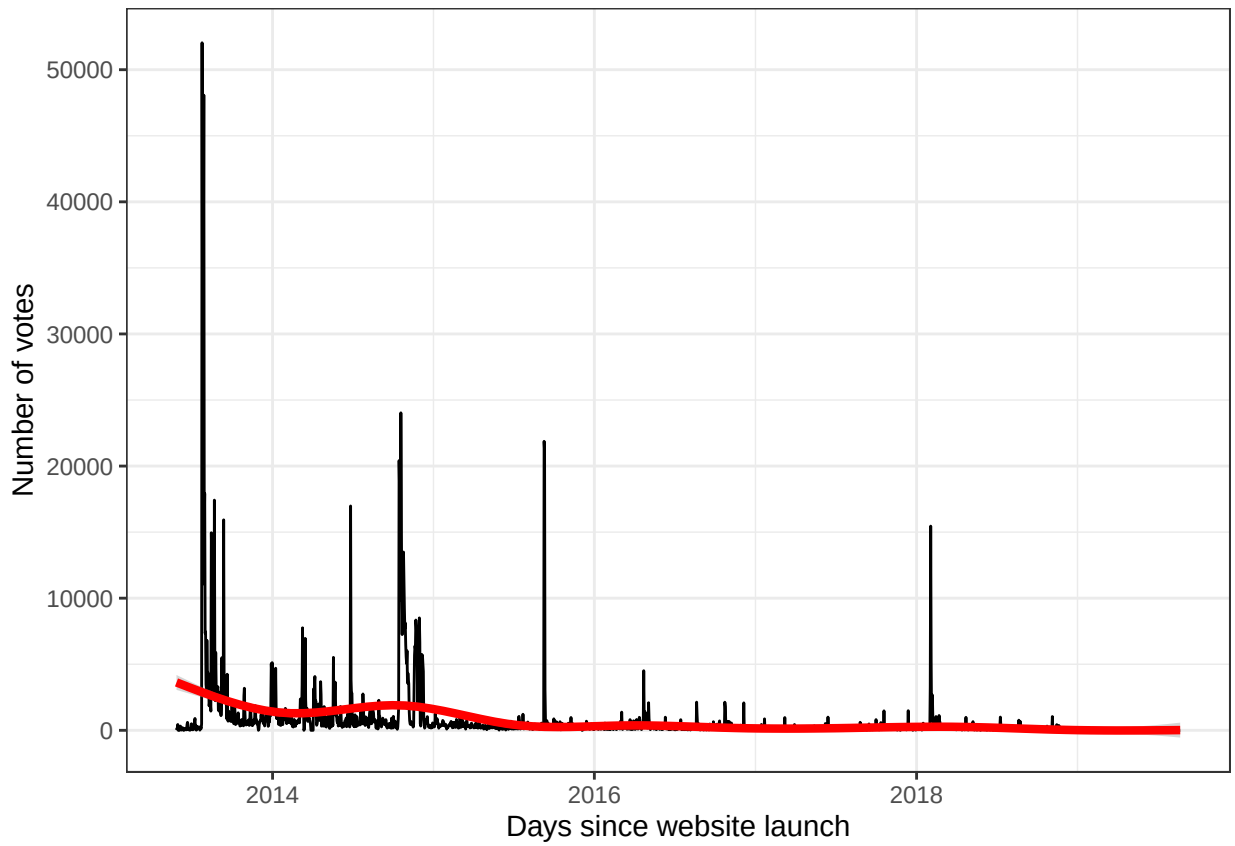
Talk a little about the issues with this and then discuss this: Note: Buil-Gil et al. (2020) Buil-Gil, Solymosi, and Moretti (2020) propose a new method to compute estimates for areas with small sample sizes

3.5.4 Study participation descrease in Place Pulse

!!!This should also be broken up and motivated/exaplained!!!

```
by_day <- pp_data %>%
          mutate(day = ymd(day)) %>%
          group_by(day) %>%
          summarise(num_votes = n()) %>%
          complete(day = seq.Date(min(day), max(day), by = "day")) %>%
          mutate(num_votes = replace_na(num_votes, 0))


ggplot(by_day, aes(x = day, y = num_votes)) +
       geom_line() +
       geom_smooth(lwd = 1.5, col = "red") +
       theme_bw() +
       xlab("Days since website launch") +
       ylab("Number of votes")
```



Note: large peak beginning on July 24th 2013: publication of Salesses, Schechtner, and Hidalgo (2013) on July 24th and press release via MIT News (http://news.mit.edu/2013/quantifying-urban-perceptions-0724)
Note: large peak beginning on October 15th 2014: publication of Harvey (2014) MSc thesis

3.4.5 Measure validity

Validity of an assessment is the degree to which it measures what it is supposed to measure.

Conclusions

Open topics in crowdsourced data

Further applications to crime and place research

Authors bios

David Buil-Gil is a Research Fellow at the Department of Criminology of the University of Manchester, UK, and a member of the Cathie Marsh Institute for Social Research at this same university. His research interests cover small area estimation applications in criminology, environmental criminology, crime mapping, emotions about crime, crime reporting, new methods for data collection and open data.

Reka Solymosi is a Lecturer in Quantiative Methods at the Department of Criminology of the University of Manchetser, UK, with interests in data analysis and visualization, crowdsourcing, rstats, fear of crime, transport, and collecting data about everyday life. As a former crime analyst, she is interested in practical applications to research and solving everyday problems with data.

References

Birenboim, A. 2016. "New Approaches to the Study of Tourist Experiences in Time and Space." *Tourism Geographies* 18 (1).

Blom, J., D. Viswanathan, J. Go, M. Spasojevic, K. Acharya, and R. Ahonius. 2010. "Fear and the City - Role of Mobile Services in Harnessing Safety and Security in Urban Contexts." Association for Computing Machinery.

Brabham, D. C. 2008. "Crowdsourcing as a Model for Problem Solving: An Introduction and Cases." *Convergence: The International Journal of Research into New Media Technologies* 14 (1).

Buil-Gil, D., R. Solymosi, and A. Moretti. 2020. "Non-Parametric Bootstrap and Small Area Estimation to Mitigate Bias in Crowdsourced Data: Simulation Study and Application to Perceived Safety." In *Big Data Meets Survey Science*, edited by C. Hill, P. Biemer, T. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, and Lyberg L. John Wiley & Sons Ltd.

Chataway, M. L., T. C. Hart, R. Coomber, and C. Bond. 2017. "The Geography of Crime Fear: A Pilot Study Exploring Event-Based Perceptions of Risk Using Mobile Technology." *Applied Geography* 86.

Elliott, M. R., and R. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32 (2).

Erete, S., L. Nicole, J. Mumm, A. Boussayoud, and I. F. Ogbonnaya-Obguru. 2016. ""That Neighborhood Is Sketchy!": Examining Online Conversations About Social Disorder in Transitioning Neighborhoods." Association for Computing Machinery.

Gabriel, U., and W. Greve. 2003. "The Psychology of Fear of Crime. Conceptual and Methodological Perspectives." *British Journal of Criminology* 43.

Gastwirth, J. L. 1972. "The Estimation of the Lorenz Curve and Gini Index." *The Review of Economics and Statistics* 54 (3).

Goodchild, M. F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69.

Goodchild, M. F., and J. A. Glennon. 2010. "Crowdsourcing Geographic Information for Disaster Response: A Research Frontier." *International Journal of Digital Earth* 3 (3).

Gómez, F., A. Torres, J. Galvis, J. Camargo, and O. Martínez. 2016. "Hotspot Mapping for Perception of Security." IEEE.

Hale, C. 1996. "Fear of Crime: A Review of the Literature." *International Review of Victimology* 4 (2).

Hamilton, M., F. Salim, E. Cheng, and S. L. Choy. 2011. "Transafe: A Crowdsourced Mobile Platform for Crime and Safety Perception Management." IEEE.

Harvey, C. W. 2014. "Measuring Streetscape Design for Livability Using Spatial Data and Methods." Master's thesis, The Faculty of the Graduate College, The University of Vermont.

Howe, J. 2006. "The Rise of Crowdsourcing." *Wired Magazine* 14 (6).

McNulty, T. L., and S. R. Holloway. 2000. "Race, Crime, and Public Housing in Atlanta: Testing a Conditional Effect Hypothesis." *Social Forces* 79 (2).

Miró-Llinares, F., A. Moneva, and M. Esteve. 2018. "Hate Is in the Air! But Where? Introducing an Algorithm to Detect Hate Speech in Digital Microenvironments." *Crime Science* 7 (15).

Pebesma, E. 2020. *Sf: Simple Features for R.* https://CRAN.R-project.org/package=sf.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Salesses, P., K. Schechtner, and C. A. Hidalgo. 2013. "The Collaborative Image of the City: Mapping the Inequality of Urban Perception." *PloS One* 8 (7).

Solymosi, R., and K. Bowers. 2018. "The Role of Innovative Data Collection Methods in Advancing Criminological Understanding." In *The Oxford Handbook of Environmental Criminology*, edited by G. J.N. Bruinsma and S. D. Johnson, 210–37. New York: Oxford University Press.

Solymosi, R., K. Bowers, and T. Fujiyama. 2015. "Mapping Fear of Crime as a Context-dependent Everyday Experience That Varies in Space and Time." *Legal and Criminological Psychology* 20 (2).

Solymosi, R., K. J. Bowers, and T. Fujiyama. 2017. "Crowdsourcing Subjective Perceptions of Neighbourhood Disorder: Interpreting Bias in Open Data." *British Journal of Criminology* 58 (4).

Solymosi, R., D. Buil-Gil, L. Vozmediano, and I. Guedes. 2020. "Towards a Place-Based Measure of Fear of Crime: A Systematic Review of App-Based and Crowdsourcing Approaches." *Environment & Behavior*.

Tester, G., E. Ruel, A. Anderson, D. C. Reitzes, and D. Oakley. 2011. "Sense of Place Among Atlanta Public Housing Residents." *Journal of Urban Health: Bulletin of the New York Academy of Medicine* 88 (3).

Warr, M. 2000. "Fear of Crime in the United States: Avenues for Research and Policy." *Criminal Justice* 4 (4).

Wickham, H., R. François, L. Henry, and K. Müller. 2020. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Williams, M. L., P. Burnap, A. Javed, H. Liu, and S. Ozalp. 2020. "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime." *British Journal of Criminology* 60 (1).

Williams, M. L., P. Burnap, and L. Sloan. 2017. "Crime Sensing with Big Data: The Affordances and Limitations of Using Open-Source Communications to Estimate Crime Patterns." *British Journal of Criminology* 57.

Zeileis, A., and C. Kleiber. 2015. *Ineq: Measuring Inequality, Concentration, and Poverty.* https://CRAN.R-project.org/package=ineq.