Using Crowdsourced Data to Study Crime and Place

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

02/05/2020

Abstract

Crowdsourcing refers to enlisting a large number of people (the crowd) through some digital platform to collect data or contribute knowledge towards a collaborative project. Crowdsourcing can generate large volumes of data in relatively little time at a very small cost, and can be useful for research, for strategic police management, and for many other purposes. To make effective use of crowdsourced data it is important to understand its key strengths to emphasise, and limitations to mitigate. In this chapter we highlight the main strengths and weaknesses, and illustrate how to acquire, make sense of, and critically evaluate crowdsourced data. We present a step-by-step exemplar study using crowdsourced data from a platform called Place Pulse, where people rate their feelings of safety between different areas. Taking the case study of Atlanta, Georgia, we work through analysing and interpreting these data while highlighing how to emphasise strengths and evaluate the limitations. Exercises are presented using `R` software.

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

Full reference: Buil-Gil, D., & Solymosi, R. (2020). Using Crowdsourced Data to Study Crime and Place. 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.

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 online technologies, mobile phones, social media or a website (Howe, 2006). Digital platforms allow recording large volumes of data in relatively little time at a small cost. Such data is often utilized for a variety of functions ranging from academic research to policy making and emergency management (Brabham, 2008; Goodchild, 2007; Hecker et al., 2019).

For 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 & Glennon, 2010). Specific to research in crime and place, crowdsourcing projects have been used to understand people's experiences with crime and their perceptions about space and safety (e.g., Solymosi & Bowers, 2018; Williams et al., 2017).

In this chapter we present some examples of how you can creatively use crowdsourcing for criminological research, and specifically highlight the strenghts and limitations of data produced from crowdsourcing platforms. We present a step-by-step exemplar study in `R` software (R Core Team, 2020) using crowdsourced perceptions of safety in Atlanta, Georgia.

Crowdsourcing and Crime and Place

In criminological research, crowdsourcing 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. It provides new angles of insight into people's activities, thus allowing researchers to devise new explanations of crime and perceived safety.

Public perceptions and emotions about crime have traditionally been analyzed by using surveys and interview-type qualitative approaches (see Gabriel & Greve, 2003; Warr, 2000), but these methods are costly and may be limited in their ability to capture the time and context-specific emotional reactions of fear. They may also fail to record any behavioral responses to such emotions, such as avoiding certain places or situations, or acquiring alarm systems or weapons. As an alternartive, some researchers have endorsed the use crowdsourcing to record data about the specific places and times in which episodes of fear of crime are more frequent (Solymosi et al., 2020; Solymosi & Bowers, 2018).

For example, Hamilton et al. (2011) developed a mobile phone app to record public perceptions of crime on public transportation in Melbourne, Australia. Similarly, Solymosi et al. (2015) designed an app and asked participants to report their worry about crime, which allowed authors to map the users' fear of crime across different areas of London, UK. 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 Bogotá, Colombia, to report those areas in which they feel less safe.

People can report not only their perceived emotions, but also about things they see in their environments. For example Solymosi et al. (2017) analyzed secondary data recorded from FixMyStreet, an online problem-reporting website, where citizens can report graffiti, broken street lights, and other signals of neighborhood disorder in London, UK.

Concerned with the effects of the built environment on people's perceptions of crime, Salesses et al. (2013) designed a website which presented people with two images from Google Street View, and asked them to choose 'which place looks safer'. This platform is called "Place Pulse", and has received thousands of views, with people all over the world evaluating images of places based on their feelings of safety. Then, based on these evaluations, Salesses et al. (2013) produced a map of perceived safety in New York. It is this specific project from which we will be analysing data later in this chapter.

These are only a few examples, but there are many other crowdsourcing platforms that have been designed and utilized to study the fear of crime (see a review in Solymosi et al., 2020).

Strengths and weaknesses of crowdsourced data

Crowdsourced data about public perceptions of space and safety have some key strengths over data recorded from traditional survey methods. Due to the data being provided by people in real-time, using technology which can record auxiliary information such as GPS or time-stamp, besides the information people report we also get precise spatial data, information about immediate environmental variables, and other relevant information, without any additional cost to researcher or participant.

However, the mode of production of crowdsourcing is also associated with certain limitations or weaknesses that, if uncontrolled, may affect the validity of such measures and the reliability and generalisability of our results (Buil-Gil et al., 2020; Elliott & Valliant, 2017).

Solymosi et al. (2020) conducted a systematic review of 27 studies utilizing or discussing the use of crowdsourcing to study perceptions and emotions about crime. Here we will summarise the key strengths and weaknesses identified in their review.

Strengths

The most frequent strength of crowdsourcing and app-based methods identified by researchers was that these techniques allow capturing the spatial-temporal specific nature of fear of crime. Unline traditional survey instruments, crowdsourcing data collection generates point-level location data, and accurate-to-the-second time-stamp data with each report. This is very beneficial for anyone carrying our crime and place research!

Another strength relevant to place-and-crime-research is the ability for people to record recording data about the architechtural features and environmental characteristics of spaces where they report (Chataway et al., 2017; Traunmueller et al., 2015). Users can provide photos of their environments (eg: with FixMyStreet (Solymosi et al., 2017)) can be asked to evaluate photos (eg: with Place Pulse (Salesses et al., 2013)) or such data can be linked from other sources by a common information, like GPS location (eg: ***DAVID STUDY***).

Crowdsourcing can also produce large sample sizes, often at a very low cost. Dubey et al. (2016), for example, analyzed more than 350,000 votes of perceived safety recorded from the Place Pulse platform; and Solymosi et al. (2017) analyzed more than 275,000 reports of disorder in London. These large samples are very costly to record by using traditional probability surveys. In this chapter we will illustrate how to download data about more than 1.5 million votes registered from the Place Pulse platform, and we will analyze more than 37,000 votes of perceived safety in Atlanta.

Limitations

Perhaps the main weakness of data recorded from crowdsourcing is related to participants' self-selection. Probability surveys are carefully designed to select participants randomly, which means that all units in the population have equal probabilities of being chosen; whereas crowdsourcing projects harness data from non-probability samples who decide when and where to share their perceptions and emotions, and whether they want to participate at all (Elliott & Valliant, 2017). The mode of production of crowdsourced data increases the risk of self-selection bias, and as a consequence males and young citizens tend to be overrepresented in these data (Chataway et al., 2017), and citizens from deprived areas are generally less represented than persons from wealthy neighborhoods (Solymosi & Bowers, 2018). For example, Salesses et al. (2013) observed that 78.3% of participants who informed about their gender when using the Place Pulse platform were males, and Solymosi et al. (2017) highlight that only 26% of those who informed about their gender when reporting instances of disorder via FixMyStreet were females.

Even within the sample, there is unequal participation. Many crowdsourcing platforms allow participants to submit data multiple times. This may lead to participation inequality (or unequal participation), where "…few users are responsible for most crowdsourced information, while the majority participate only a few times" (Buil-Gil et al., 2020, p. 6). To illustrate this, Dubey et al. (2016) show that 6,118 of the 81,730 persons who used the Place Pulse platform participated only once, while 30 users participated more than 1,000 times and the most prolific user voted 7,168 times. Solymosi et al. (2017) also show that one fourth of all FixMyStreet reports are produced by one percent of participants, and 73% of participants contribute only once.

It is not only people who may not all be represented in these sorts of data, there may also be a bias in the places and times that do or do not feature prominently. Since users of crowdsourcing projects can decide where and when to participate means that certain types of areas and times can be underrepresented. For instance, it is possible that app-based platforms fail to capture data from high-crime-density areas, since participants may avoid those places where they feel more exposed to crime (Innes, 2015). The routine activities of participants are also reflected on an under-representation of data points at night (Blom et al., 2010) when people are less likely to be out and about.

Also, many of these projects rely from enthusiasm from the crowd of participants, which may die down over time. It is important to consider when such a project is launched, as the beginning is much more likely to see lots of active participation than later on in the project.

Finally, there are ethical considerations that may arise from the use of crowdsourcing, which are related to the user's privacy (e.g., risk of participants' identification) but also concerns that these techniques may sensitize participant and increase their fear of crime by asking them to constantly think about crime-related risks (Jackson & Gouseti, 2015; Solymosi et al., 2020). This last point is not so much a limitation as a key concern that all reasearchers working with crowdsourced data should keep at the forefront of their minds.

Summary

Overall crowdsourced data can be understood as data provided by contributions from a large crowd of people through some digital platform to collect data or contribute knowledge towards one collaborative project.

Key strength of these data include:

- precise spatial and temporal information
- ability to collect contextual information about the environment
- large sample sizes often at low cost

Key limitations to always address with such data include:

- Issues with generalisability from non-representative samples
- Unequal contribution from individuals within the sample
- Possible unequal geographical or temporal coverage
- Initial enthusiasm of participation will die down

It is further important to think about ethical concerns around the right to privacy of those contributing to these types of crowdsourcing platforms.

Crowdsourcing perceptions of safety: A step-by-step example in `R`

In order to illustrate the use of crowdsourcing in criminological research, we present an exemplar study using data recorded by the Place Pulse 2.0 platform (Salesses et al., 2013) mentioned earlier. This section will introduce the Place Pulse project and provide annotated `R` scripts for downloading, exploreing and cleaning this source of crowdsourced data. Then, we will analyze the spatial distribution of crowdsourced perceptions of space and safety in Atlanta and illustrate with examples how to explore some of the known issues of crowdsourced data discussed above in a real-world dataset.

The Place Pulse project

Place Pulse 2.0 was an online crowdsourcing platform designed to record data about citizens' perceptions of a variety of topics including safety, beauty, wealth, liveability, boredom and depression in urban areas. Each topic had a related question. Here we will focus onsafety. To assess safety, two images were shown to participants, who then were asked to answer *'Which place looks safer?'* (see Figure 1). Participants could also be asked which of the two images looked wealthier, more beautiful, more boring, livelier or more depressing, but we will focus on perceptions of space and safety in this chapter.

The images were selected randomly from Google Street View images of 56 cities across 28 countries. The images were all taken between 2007 and 2012. All the data collected were stored on the Place Pulse open website (http://pulse.media.mit.edu/), and available to everyone there, but the platform closed in late 2019. We have been granted access to all the data recorded between May 28th 2013 and August 22nd 2019 to write this chapter. All data that have also been uploaded onto an open repository with consent of the data producers (Salesses et al., 2013).

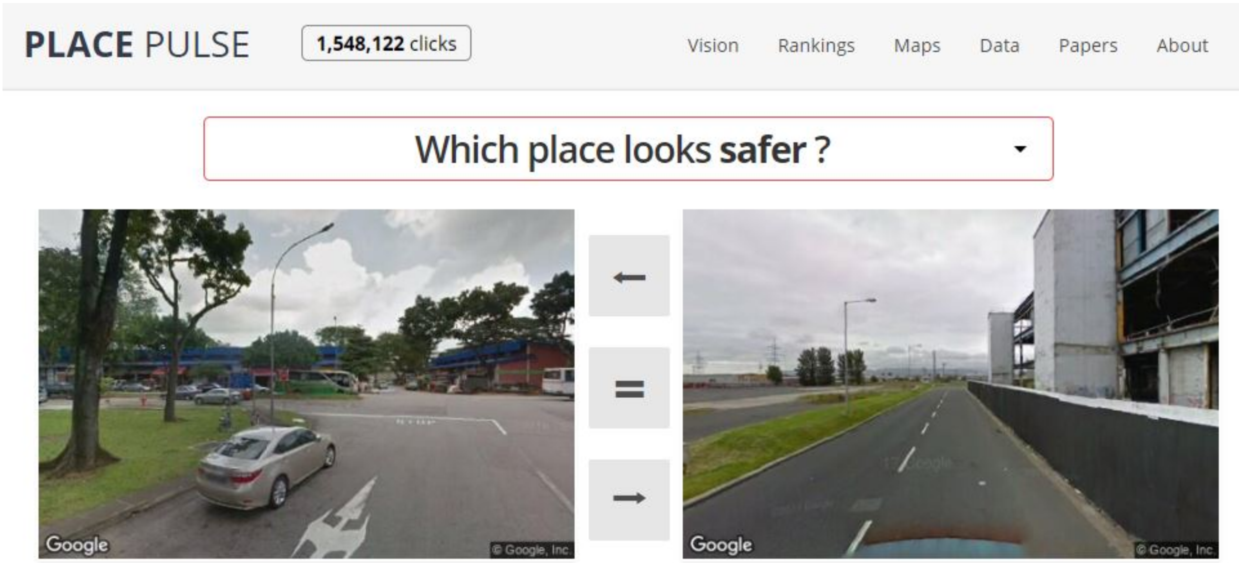


Figure 1: Figure 1: Place Pulse website

Download and explore Place Pulse data

We have saved all Place Pulse data (more than 1.5 million votes) in a data repository on FigShare. You can download this directly into `R` by using the `read.csv()` function.

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

This dataset includes 17 variables, but we will only use some of them, so let's use the `select()` function from the `dplyr` package to select only the ones we need. We will need the following variables:

- *'X':* A unique identification code for each vote.
- *'right':* A unique identification code for the image in the right side of the pairwise comparison.
- *'left':* A unique identification code for the image in the left side of the pairwise comparison.
- *'voter_uniqueid':* A unique identification code given to each participant.
- *'place_id_left':* A unique identification code for the place in the image in the left side of the pairwise comparison.
- *'place_id_right':* A unique identification code for the place in the image in the right side of the pairwise comparison.
- *'place_name_left':* The name of the city in the left side of the comparison.
- *'place_name_right:'*The name of the city in the right side of the comparison.
- *'choice':* Which image was selected as 'safet (left, right, or equal) .
- *'study_question':* The variable of interest (e.g., safety, wealth, beauty).
- *'day':* Day of the data point (vote)
- *'time':* Day of the data point (vote)
- *'long_right':* Longitude for image on the right side of the comparison.
- *'lat_right':* Latitude for image on the right side of the comparison.
- *'long_left':* Longitude for image on the left side of the comparison.
- *'lat_left':* Latitude for image on the left side of the comparison.

```
library(dplyr)

pp_data <- select(pp_data, X, left, right, voter_uniqueid, place_id_left, place_id_right, place_name_le
```

We can also see we have 1565723 observations (that's a lot!). Each observation is one comparison between two images. One vote by a participant. So our unit of analysis here, is each comparison, between an image on the left and an image on the right.

We can start by answering some descriptive questions, for example, which cities are most frequenty assessed within the Place Pulse platform?

We know that for each row (each comparison) there are two images (`place_id_left` and `place_id_right`), and there are two columns that name the location for each image (`place_name_left`, `place_name_right`). To see how many times each image appears we can create two frequency tables (one for `place_name_left`, and one for `place_name_right`), and then join them, and sum the two frequencies. We can use the `group_by()` and `summarize()` functions from `dplyr` package (Wickham, François, et al., 2020) to achieve this:

```
right_freq <- pp_data %>%
  group_by(place_name_right) %>% # categories based on cities on the left
  summarize(right_count = n())

left_freq <- pp_data %>%
  group_by(place_name_left) %>% # categories based on cities on the right
  summarize(left_count = n()) # count number of units in each category

total_freq <- left_freq %>%
  left_join(., right_freq, by = c("place_name_left" = "place_name_right")) %>%
  mutate(total_count = left_count + right_count)
```

Now we can see the top 3 most common cities using the `top_n()` function:

```
total_freq %>% top_n(3)
```

```
## # A tibble: 3 x 4
##   place_name_left left_count right_count total_count
##   <fct>                <int>       <int>       <int>
## 1 Atlanta              56992       57140      114132
## 2 Berlin               55265       55583      110848
## 3 Tokyo                53817       53514      107331
```

Atlanta appears to be 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))  %>% # print in descending order
  filter(!is.na(study_question)) #remove NAs
```

```
## # A tibble: 6 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
```

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 analyze if responses were biased by the position in which images were shown on the website.

This is important to do because it confirms the validity of this study instrument. In survey design, much thought and research goes into elements like the ordering of the questions. It is important that we consider crowdsourced platforms with the same care and attention.

```
pp_data %>%
  group_by(choice) %>% # categories based on vote (right, left or equal)
  summarize(count = n()) %>% # count number of units in each category
  mutate(`%` = round(count/sum(count), 3)*100) %>%
  top_n(3) # print 3 most frequent categories
```

```
## # A tibble: 3 x 3
##   choice  count   `%`
##   <fct>   <int> <dbl>
## 1 equal  206147  13.2
## 2 left   668680  42.7
## 3 right  690792  44.1
```

```
#proportion
```

The frequency of votes for left and right options is very similar; the left image wins about 44% of the time, the right one about 43% of the time, and the rest of the time it's a tie (13%). Thus we can conclude that the position of the image on the website platform does not appear to have much of an affect on participants' votes.

Cleaning Place Pulse data

Our city of interest, Atlanta is the capital city of the State of Georgia, United States. In 2018, its estimated papulation was close to 500,000 residents, and it is the 37th most populated city in the United States. It is also, as we have seen earlier, the city with the largest number of votes in the Place Pulse platform. There has been some considerable research by criminologists looking into predictors of crime and fear of crime in this city, which can be used to interpret our findings later (see McNulty & Holloway, 2000; Tester et al., 2011)

So to find out about perception of safety in Atlanta using our crowdsourced Place Pulse data, the first step is to clean the data to make it as complete and useful as possible to answer our research questions. For example, in this case, we want to map the perceived safety of areas in Atlanta only, however we have reports from all over the world. Also, we would like our unit of analysis to be the Atlanta locations, rather than each comparison. You can see, there are a few steps that we need to take to make the data look like what we need to answer our questions. Specifically we need to do the following:

- Select only votes about safety (vs beauty etc)
- Select only those votes that contain images of Atlanta
- For each image in Atlanta, create a score from all the votes

Select only votes about safety

We can use the function `filter()` from `dplyr` to create a new dataframe that includes those pairwise comparisons which relate to safety:

```r
pp_s <- pp_data %>%
  filter(study_question == "safer") # select votes of 'safer'
```

This has created a new dataframe, `pp_s` which contains all 509961 safer votes.

Select only votes that contain images of Atlanta

Now we want to select those votes where at least one of the images is from Atlanta. Remember there are two images in each comparison `place_name_right` and `place_name_left`. In order to select rows where Atlanta appears in either one or the other, we can use the or operator (`|`) to ensure we get all comparisons which feature Atlanta on at least one side.

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

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

Reshape the dataset so each time a place in Atlanta has lost or won a vote is our unit of analysis

You may notice, that the dataset we want to answer our research question is slightly different to the dataset we have from the crowdsourced platform. This is often the case with much secondary data analysis, where data were created for another purpose initially. This often means that much data wrangling is to be done to make the data look like what you need.

We are interested in analyzing the proportion of 'safer' votes in each neighborhood of Atlanta. For this we need for each photo in Atlanta two things: 1) it's location and 2) whether it won or lost a vote.

So first we need every instance of a vote on each unique location in Atlanta. There are three possibilities for this:

1) Image on the right is Atlanta but image on the left is somewhere else.
2) Image on the left is Atlanta, but image on the right is somewhere else.
3) Both images are in Atlanta.

For the first two cases, we can easily find the coordinates of the image in Atlanta, and whether it lost or not, using conditional statements in the `if_else()` function.

Let's first create a dataset of only these votes. Further, in order to assign photographs to their neighborhood, we need to create two new columns that specify the longitude and latitude of each image of Atlanta being assessed. We will also create a new column that details whether each participant voted that the image of Atlanta was 'safer' than the other photograph. Some pairwise comparisons, however, assessed two different images from Atlanta, which means that we will need to duplicate those votes to count both images. 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 also used above, but now, instead of the or operator (`|`) we use an *and* operator (`&`). We also compute the win score, and save a unique ID for each image:

```r
atl_v_others <- pp_atl_s %>%
    filter((place_name_right == "Atlanta" & place_name_left != "Atlanta") |    # select where only one s
             (place_name_right != "Atlanta" & place_name_left == "Atlanta")) %>%
  mutate(lat_Atl = if_else(place_name_right == "Atlanta", lat_right, lat_left), # get latitude from Atl
```

```r
        long_Atl = if_else(place_name_right == "Atlanta", long_right, long_left), # get longitude from
        win = if_else((place_name_left  == "Atlanta" & choice == "left") |
                      (place_name_right == "Atlanta" & choice == "right"), 1, 0),  # label Atlanta sid
       unique_id = if_else(place_name_right == "Atlanta", right, left))
```

For the last case, where both images are of Atlanta, we need to separate out a new line for each image. For this we can create a dataset for the right side and one for the left side, also creating the new latitude (`lat_Atl`), longitude (`long_Atl`), and win (`win`) columns for each using `mutate()`:

```r
#create dataset from Atlanta images on right side
right_side <- pp_atl_s %>%
    filter(place_name_right == "Atlanta" & place_name_left == "Atlanta") %>%
  mutate(lat_Atl = lat_right,
         long_Atl = long_right,
         win = if_else(choice == "right", 1, 0),
         unique_id = right)

#create dataset from Atlanta images on left side
left_side <- pp_atl_s %>%
    filter(place_name_right == "Atlanta" & place_name_left == "Atlanta") %>%
  mutate(lat_Atl = lat_left,
         long_Atl = long_left,
         win = if_else(choice == "left", 1, 0),
         unique_id = left)
```

Finally, we can merge all the data together with the `rbind()` function:

```r
pp_atl_s <- rbind(atl_v_others, left_side, right_side)
```

We have a dataframe of 37892 votes about the safety in Atlanta that is ready to be analyzed, which now fits our criteria, in that:

- it contains only votes about safety
- it contains only votes about Atlanta
- each row is a win (or no win) of a vote where an image from Atlanta was chosen as safer (or not safer)

We can now take this dataset and draw some conclusions.

Environmental correlates of safety

Researchers may be interested in analyzing the environmental characteristics of those places assessed by Place Pulse users as safe or not safe.

From such data we can get the coordinates of places rated, and use tools such as Google Street view to conduct a virtual environmental audit of these places. For example, let's look at the safest and least safe rated places using our dataset in Atlanta.

For this, we need a dataset where each unique image is our unit of analysis, and we can conpute a win score by considering the proportion of votes which that image has won:

```r
pp_atl_s %>%
  group_by(unique_id) %>%
  summarize(winscore = mean(win, na.rm = TRUE), # calculate proportion of 'safer'
            num_votes = n(),  # count number of votes
            latitude = first(lat_Atl), longitude = first(long_Atl)) %>% # save latitude and longitude
  filter(num_votes >= 20) %>% # filter out pictures with less than 20 votes
  arrange(desc(winscore)) %>% # order by proportion of 'safer' responses
  filter(row_number() == 1 | row_number() == n()) # print lowest and highest score
```

```
## # A tibble: 2 x 5
##   unique_id                winscore num_votes latitude longitude
##   <fct>                       <dbl>     <int>    <dbl>     <dbl>
## 1 513d7c40fdc9f03587006e2b    0.864        22    -84.4      33.7
## 2 513d7a35fdc9f03587006757    0.05         20    -84.3      33.8
```

This may be used by criminologists to observe the characteristics of those places assessed by Place Pulse participants as more or less safe. For example, we can use Google Street View [https://www.google.com/maps] to observe places rated as the least safe or safest amongst those rated at least 20 times (see Figure 2). In this case, the least safe place is characterized by very dense vegetation (which may be perceived to offer concealment for possible criminals and obstructs the view onto the ground), lack of exit routes to escape from potential threats, and an abandoned house with signs of physical disorder (see Fisher & Nasar, 1992); whereas the safest place is a wider street of a residential area with direct visual access to most places around it (large prospect), lack of places for concealment of offenders and natural surveillance from houses (Welsh & Farrington, 2004). We can do much more, but here we will focus on the specific issues to explore due to the crowdsourced nature of these data.

Mapping Place Pulse data

Now we can use mapping techniques learned in other chapters (LINK WITH MAPPING CHAPTERS) to map crowdsourced data. We will be using the `sf` (Pebesma, 2020) and `ggplot2` (Wickham, Chang, et al., 2020) libraries in order to create a map of perceived safety of built environment across the areas of Atlanta.

First, acquire a polygon which represents Atlanta. The Georgia Association of Regional Commission, for instance, publishes spatial data for Atlanta Region at the different spatial scales. You can go on their website to find out more about this boundary data:%5Bhttps://opendata.atlantaregional.com/datasets/census-2000-tracts-atlanta-region](https://opendata.atlantaregional.com/datasets/census-2000-tracts-atlanta-region). We can download the shapefile directly using their Application Programme Interface (or API) and the `st_read()` function form the `sf` package:

```r
library(sf)
# download geojson from Georgia Association of Regional Commissions open data
atl <- st_read("https://opendata.arcgis.com/datasets/04b79404794f43959cda4f8c3f1817e6_49.geojson")
```

```
## Reading layer `cdd56316-b32c-4dfc-96c6-d770313b03ab2020411-1-l4zst3.e884q' from data source `https:/
## Simple feature collection with 712 features and 21 fields
## geometry type:  POLYGON
## dimension:      XY
## bbox:           xmin: -85.38659 ymin: 32.84465 xmax: -83.50573 ymax: 34.61792
## CRS:            4326
```

(note if the link doesn't work, we saved a local copy of this file, in that case you can download from the following link: https://raw.githubusercontent.com/maczokni/crowdsourcing__pp__chapter/master/geojson/Census__2000__Tracts__Atlanta__Region.geojson)

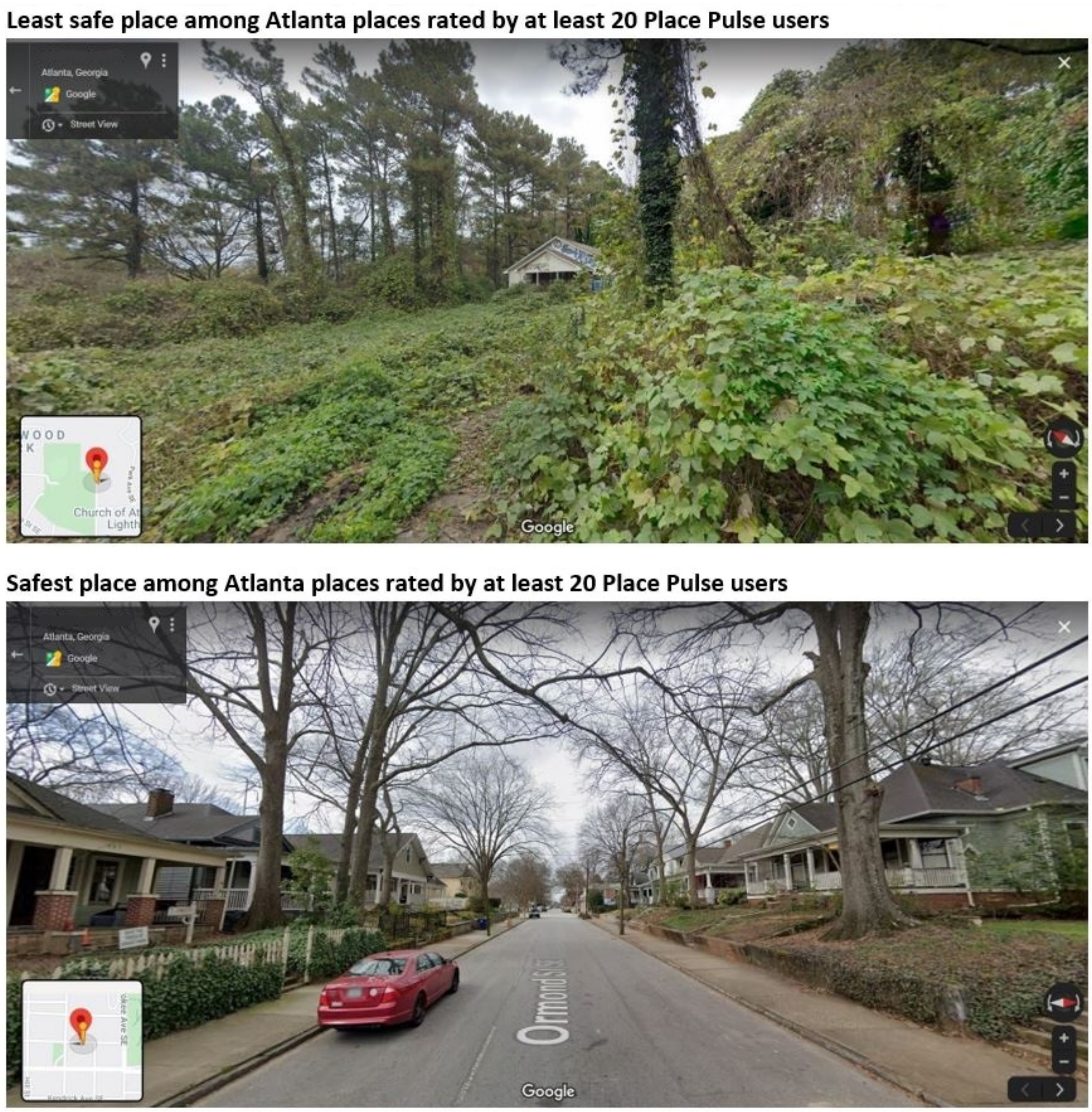
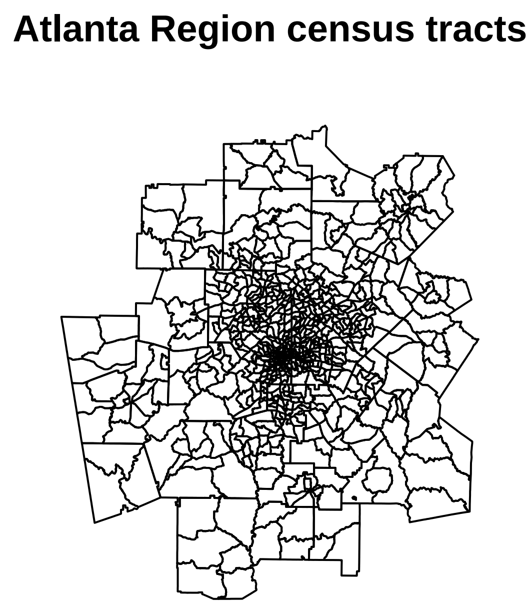


Figure 2: Figure 2: Least safe and safest places rated by at least 20 Place Pulse users

We can see what this file looks like by using the `plot()` and `st_geomerty` functions to plot the geometry of the 'atl' object we created:

```
plot(st_geometry(atl),
     main = "Atlanta Region census tracts")
```



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 assign the CRS from our polygon to our points layer:

```
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 in both layers
```

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

Now, to map ‘safer’ votes per census tract, we will compute a proportion of wins in each tract that will allow us to directly analyze the geographical distribution of perceived safety (see Buil-Gil et al., 2020).

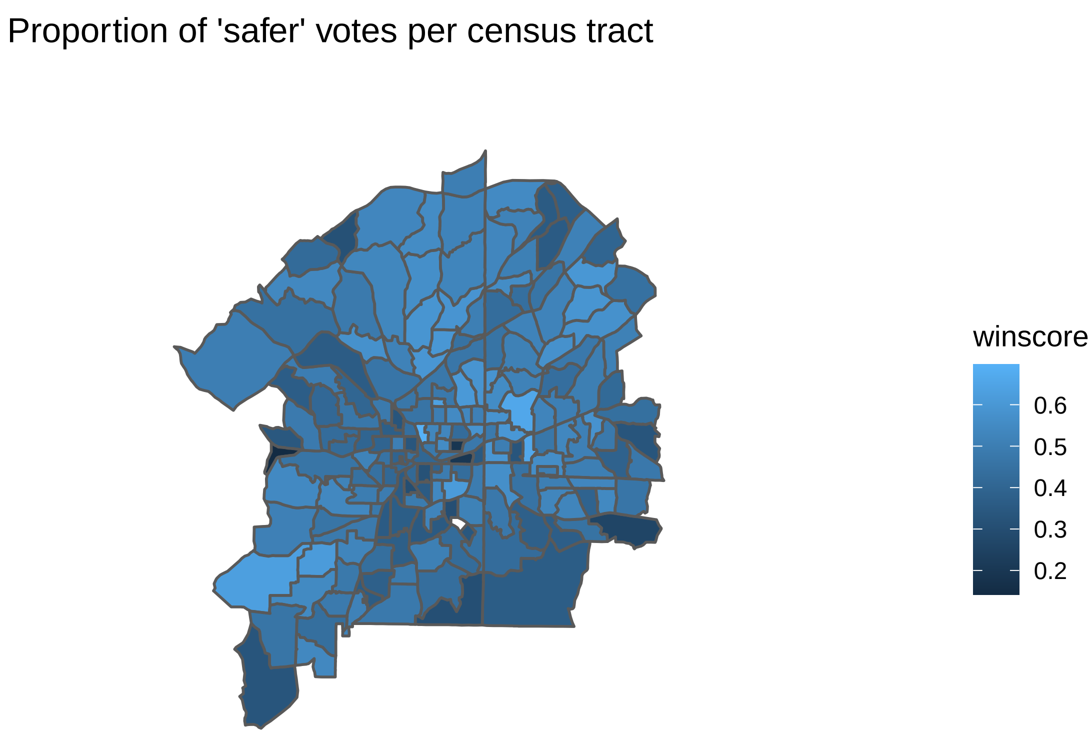
```
points_atl_s_nhood <- st_intersection(atl, points_atl_s) %>% # intersection of points and areas
                      group_by(TRACT) %>% # make groups based on tracts
                      summarise(winscore = mean(win, na.rm = TRUE), # proportion 'safer'
                                num_votes = n()) %>%  # count number of votes
                      st_set_geometry(NULL) #remove geometry (so we can join to polygon layer)
```

We now have a dataframe which has the proportion of wins (`winscore`) for each tract. All that is left is to join this to our polygon of Atlanta (`atl`) and produce a map.

```
# merge census tracts and Place Pulse votes based on common 'TRACT' column
atl_pp_wins <- left_join(atl, points_atl_s_nhood, by = c("TRACT" = "TRACT")) %>%
  filter(!is.na(winscore)) # delete census tracts with 0 votes (NAs)
```

Finally, map the propotion of ‘safer’ votes in each census tract.

```
ggplot(data = atl_pp_wins) +
  ggtitle("Proportion of 'safer' votes per census tract") +
  geom_sf(aes(fill = winscore)) +
  coord_sf(xlim = c(-84.7, -84), ylim = c(33.5, 34), expand = FALSE) +
  theme_void()
```



Exploring known issues of crowdsourced data within Place Pulse

As we set out earlier in the chapter, crowdsourced data comes with many possible issues which need to be properly understood. Here we ser out how to explore some of these issues, working through examples from Place Pulse.

How representative is the sample?

Place Pulse data contains vote from a self-selected sample, and unfortunately, 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 (Chataway et al., 2017; Elliott & Valliant, 2017).

However the sample's self-selection bias should be checked when possible. In this case, we do know that the first edition of Place Pulse did record some demographic variables from participants, and that in that iteration, 78.3% of participants were males, and only 21.7% were females, and the median self-reported age was 28 years (Salesses et al., 2013). We can consider that the version of Place Pulse we explored may have similar characteristics.

Participation inequality ('supercontributors')

Another issue is participation inequality within the sample. Crowdsourced data tend to be affected by a few number of supercontributors that produce most votes (Dubey et al., 2016; Solymosi et al., 2017). In order to check if our dataset is affected by this, we can use the variable `voter_uniqueid` and produce a frequency table:

```
voter <- pp_data %>%
  group_by(voter_uniqueid) %>% # create groups based on users unique id
  summarise(num_votes = n()) # print the number of votes by user
```

We can have a look at this new dataframe using the `View()` function, and see that we have some *very active* participants. The top participant, for instance, has made 7168 votes on places. That is some very prolific participation. On the other hand, we 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 many 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 see that this new dataframe contains 954 people, who are our top 1% contributors to the Place Pulse dataset. We will now examine how much of the total number of votes are generated by the top 1% of users:

```
# proportion of votes by top 1% participants
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. We can also compute the proportion of votes made by the top 10% and 25% of participants in the same way.

Quantifying participation inequality

One way to quantify the extent to which participation inequality exists in our data is by using a Gini index, and visualizing 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 is generally used to examine income inequality, it has also been frequently used to assess participation inequality in crowdsourcing platforms (see Solymosi et al., 2017; Solymosi & Bowers, 2018). Similarly, the Lorenz curve is a visual representation of inequality. For this we will need the `ineq` library (Zeileis & Kleiber, 2015)

Then we can load this library and calculate the index using the `Gini()` function applied to our number of votes column in our frequency table dataframe:

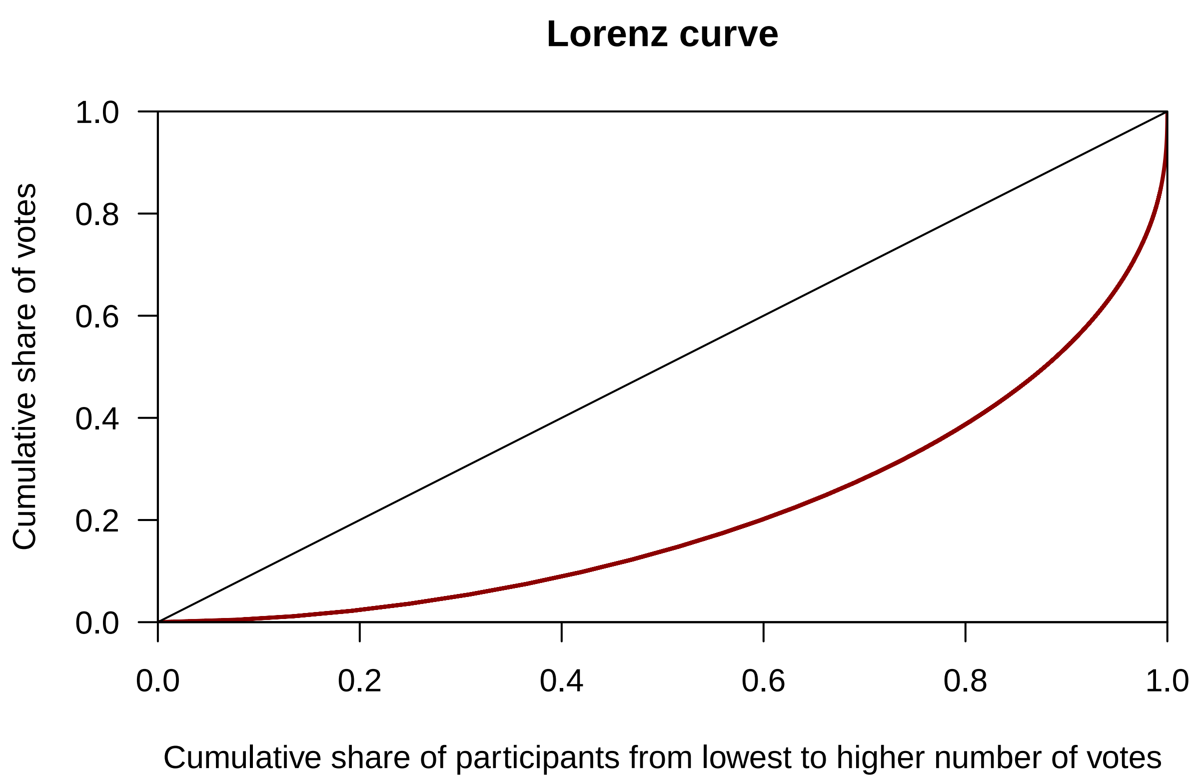
```
library(ineq)

Gini(voter$num_votes) # print Gini index
```

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

To interpret this number, we can consider the following. A Gini index score of 0 represents perfect equality (everyone makes equal number of votes), while 1 shows perfect inequality (only one person making every single vote). 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 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), # plot Lorenz curve
     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) shows how the top few percent of users 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.

Under-representation of certain areas

Another important consideration is the variation in the sample size across different areas. Some places might have many votes, while others not so much. While Place Pulse might not suffer from underreporting of scary areas due to people avoiding them (people are randomly presented with images from all over, to vote), it is still important to consider the number of votes in each census tract for example. We can analyze if certain areas are under-represented in our dataset by using the `summary()` function applied to the number of votes (`num_votes`) variable in our polygon object from earlier (`atl_pp_wins`).

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    85.0   151.0   192.9   248.0  1023.0
```

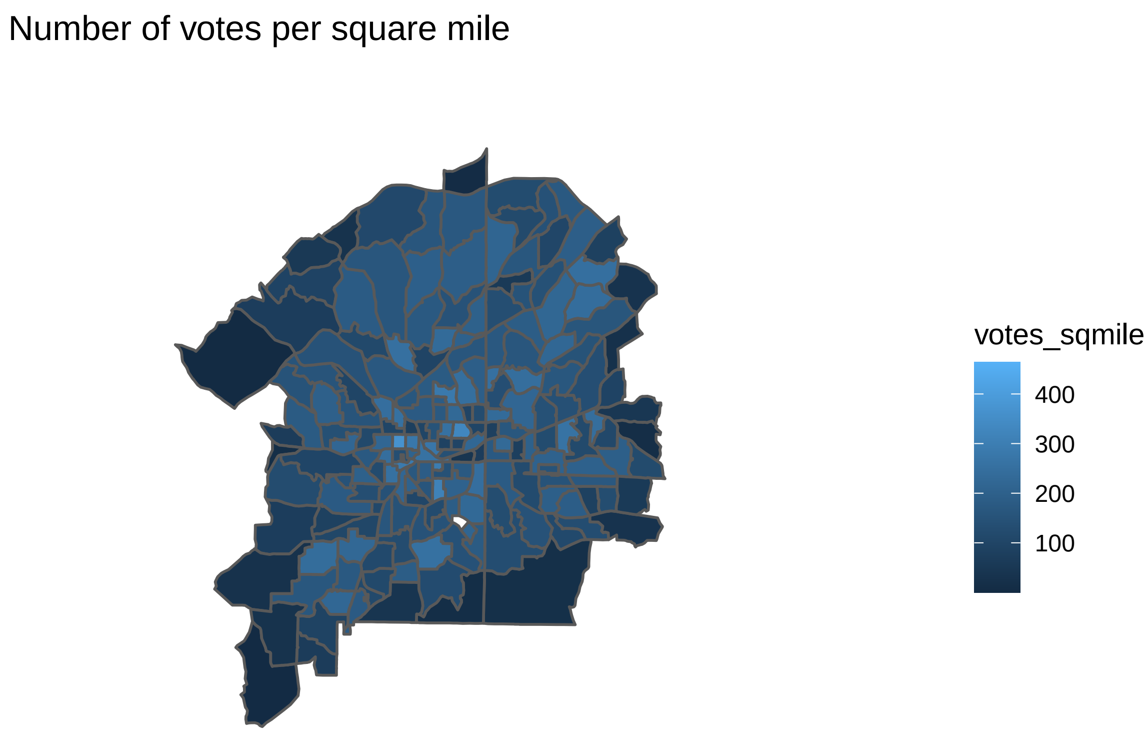
Whereas the average sample size per area is quite large (192.89), some tracts are clearly over-represented (the maximum number of votes is 1023) and others suffer from small representation (the minimum number of votes is only 2).

We may also want to know which areas suffer from severe under-representation in our dataset. We use the function `mutate()` to compute the number of votes divided by square miles in each census tract.

```
# compute new column of number of votes divded by square miles
atl_pp_wins <- atl_pp_wins %>%
  mutate(votes_sqmile = num_votes / SQ_MILES)
```

Then we can visualize the geographic distribution of the number of votes per census tract using the same code shown above to map perceptions of safety.

```
ggplot(data = atl_pp_wins) +
   ggtitle("Number of votes per square mile") +
   geom_sf(aes(fill = votes_sqmile)) +
   coord_sf(xlim = c(-84.7, -84), ylim = c(33.5, 34), expand = FALSE) +
   theme_void()
```



We see that areas in the city center tend to have larger number of votes and are therefore well represented, whereas tracts in surrounding areas suffer from smaller sample sizes. Estimates of perceived safety in under-represented areas are likely to be affected by a small number of responses and may suffer from low precision. In order to increase the reliability of estimates produced from crowdsourced data for areas with small sample sizes, some researchers suggest using resampling and model-based techniques (see Arbia et al., 2018; Buil-Gil et al., 2020).

Participation decrease

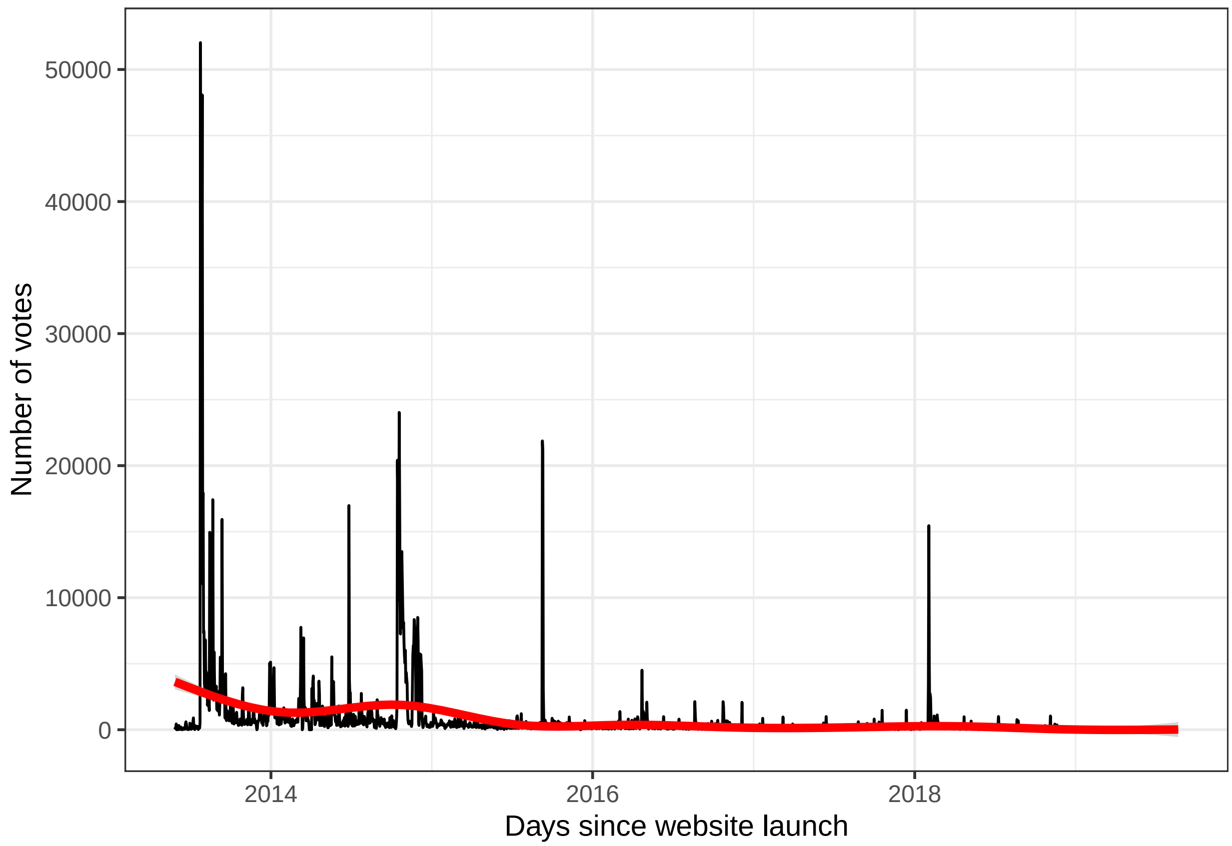
Finally, some researchers have identified that the number of users of crowdsourcing projects decreases over time: whereas the number of participants tends to be large during the first few days, users lose interest in the project if they do not obtain clear short-term benefits from using it (see Blom et al., 2010; Solymosi et al., 2020). We can also expore whether our Place Pulse dataset is affected by participation decrease.

We will use various functions from `dplyr` and `ggplot2` packages (seen above) to visualize the number of votes since the Place Pulse website launch until its closure. But we also need to use other key functions: *(i)* the `ymd()` function from `lubricate` package is used to transform the dates in which votes took place into `Date` objects (Spinu et al., 2020), and *(ii)* the `complete()` function from `tidyr` package is used to turn implicit missing dates into explicit missing dates and create a timeline of dates with and without votes (Wickham & Henry, 2020).

```
by_day <- pp_data %>%
          mutate(day = ymd(day)) %>% # create column by transforming into 'Date' object
          group_by(day) %>% # create groups by days
          summarise(num_votes = n()) %>% # count number of votes by day
          complete(day = seq.Date(min(day), max(day), by = "day")) %>% # complete all days
```

```r
         mutate(num_votes = replace_na(num_votes, 0)) # create new column: votes by day

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")
```



The number of votes within the Place Pulse platform clearly decreased over time, but we also observe some peaks even years after the launching of the project. Some of these peaks match the dates of key publications using Place Pulse data and media releases, which shows that participation in crowdsoucing projects can be enhanced by periodic campaigns. For example, we observe a large peak beginning on July 24th 2013, date in which Salesses et al. (2013) published their paper and the Massachusetts Institute of Technology published a news article about the Place Pulse platform on their website: http://news.mit.edu/2013/quantifying-urban-perceptions-0724. We also observe another peak of participation beginning on October 15th 2014, just after the publication of Harvey (2014) Master's thesis about how to automate the study of the characteristics of streetscape skeletons and urban perceptions from Place Pulse data.

Conclusions

The open data movement has provoked a revolution in social research methods, and will continue changing the way in which many social issues are researched, understood and managed. Digital technologies enable large volumes of data to become available for social researchers and data scientists, and crowdsourcing is becoming a key source of data to analyze and map social phenomena such as crime (Bendler et al., 2014) and perceptions of space and safety (Solymosi et al., 2015; Solymosi & Bowers, 2018). In this chapter we have described and explored the main strengths and weaknesses of using crowdsourced data for criminological research. Specifically, we have obtained access to a large dataset of more than 1.5 million votes about urban perceptions recorded from the Place Pulse project (Salesses et al., 2013), selected a sample of more than 37,000 votes of perceived safety for Atlanta, and studied the spatial distribution of perceptions of space and safety at a census tract level in this city. We have also shown how these data can be utilized to identify places assessed by participants as very safe or very unsafe; places in which researchers can then conduct observation to study those environmental features that make citizens feel fear of crime (Fisher & Nasar, 1992).

Although crowdsourcing offers advantages over traditional survey methods to study perceptions and emotions about crime, data recorded from crowdsourcing is also affected by certain issues that, if uncontrolled, are likely to affect the validity of data and the reliability and generalizability of research outputs. For instance, we have observed how Place Pulse votes are largely produced by a few number of super-contributors (i.e., participation inequality), there is under-representation of certain areas outside the city center, and the number of votes decreases over time (i.e., participation decrease). These issues have also been observed in data produced from many other crowdsourcing and app-based projects (e.g., Chataway et al., 2017; Solymosi et al., 2015, 2020; Traunmueller et al., 2015). Other researchers have also highlighted that crowdsourced data tends to be affected by self-selection bias, which explains why males tend to participate more than females, and young persons more than adults (Solymosi & Bowers, 2018); but the Place Pulse platform did not record demographic variables from participants and we have not directly assessed this issue here.

Due to the fact that crowdsourced datasets - and non-probability samples in general - may be affected by these potential sources of unrepresentativeness and bias, several researchers are exploring new techniques to enable obtaininig reliable research outputs. Elliott & Valliant (2017), for example, present different methods to compute individual pseudo-sampling weights and adjust non-probability samples to target populations; Arbia et al. (2018) have developed a method to delete spatial outliers and calculate weights to adjust non-probability samples to optimal spatial samples; and Buil-Gil et al. (2020) investigate the use of resampling and model-based small area estimation techniques to allow producing reliable estimates at detailed spatial scales from crowdsourced data. Academics and practitioners will benefit from methods to mitigate the sources of bias in crowdsourced data, which may allow obtaining more precise and reliable - but also cheaper - findings and devise new explanations of crime, antisocial behavior and emotions about crime. In the context of crime analysis, bias-corrected crowdsourced data may become a key tool to understand crime patterns, anticipate crime trends and even provide assistance to police investigations (Bendler et al., 2014; Nhan et al., 2017).

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.

Acknowledgments

The authors would like to thank César A. Hidalgo for providing access to the data used in this book chapter.

References

Arbia, G., Solano-Hermosilla, G., Micale, F., Nardelli, V., & Genovese, G. (2018). *Post-sampling crowd-sourced data to allow reliable statistical inference: The case of food price indices in nigeria*. Open Conference Systems.

Bendler, J., Brandt, T., Wagner, S., & Neumann, D. (2014). *Investigating crime-to-twitter relationships in urban environments - facilitating a virtual neighborhood watch*. Association for Information Systems.

Birenboim, A. (2016). New approaches to the study of tourist experiences in time and space. *Tourism Geographies*, *18*(1).

Blom, J., Viswanathan, D., Go, J., Spasojevic, M., Acharya, K., & Ahonius, R. (2010). *Fear and the city - role of mobile services in harnessing safety and security in urban contexts*. Association for Computing Machinery.

Brabham, D. (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., Solymosi, R., & Moretti, A. (2020). Non-parametric bootstrap and small area estimation to mitigate bias in crowdsourced data: Simulation study and application to perceived safety. In C. Hill, P. Biemer, T. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L.L. (Eds.), *Big data meets survey science*. John Wiley & Sons Ltd.

Chataway, M., Hart, T., Coomber, R., & Bond, C. (2017). The geography of crime fear: A pilot study exploring event-based perceptions of risk using mobile technology. *Applied Geography*, *86*.

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. (2016). Deep learning the city: Quantifying urban perception at a global scale. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision - eccv 2016* (pp. 196–212). Springer.

Elliott, M., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2).

Fisher, B., & Nasar, J. (1992). Fear of crime in relation to three exterior site features. *Environment and Behavior*, *24*(1).

Gabriel, U., & Greve, W. (2003). The psychology of fear of crime. Conceptual and methodological perspectives. *British Journal of Criminology*, *43*.

Gastwirth, J. (1972). The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, *54*(3).

Goodchild, M. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*.

Goodchild, M., & Glennon, J. (2010). Crowdsourcing geographic information for disaster response: A research frontier. *International Journal of Digital Earth*, *3*(3).

Gómez, F., Torres, A., Galvis, J., Camargo, J., & Martínez, O. (2016). *Hotspot mapping for perception of security*. IEEE.

Hamilton, M., Salim, F., Cheng, E., & Choy, S. (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.

Hecker, S., Wicke, W., Haklay, M., & Bonn, A. (2019). How does policy conceptualise citizen science? A qualitative content analysis of international policy documents. *Citizen Science: Theory and Practice*, *4*(1).

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6).

Innes, M. (2015). 'Place-ing' fear of crime. *Legal and Criminological Psychology*, *20*(2).

Jackson, J., & Gouseti, I. (2015). Psychological proximity and the construal of crime: A commentary on “mapping fear of crime as a context-dependent everyday experience that varies in space and time”. *Legal and Criminological Psychology*, *20*(2).

McNulty, T., & Holloway, S. (2000). Race, crime, and public housing in atlanta: Testing a conditional effect hypothesis. *Social Forces*, *79*(2).

Nhan, J., Huey, L., & Broll, R. (2017). Digilantism: An analysis of crowdsourcing and the boston marathon bombings. *British Journal of Criminology*, *57*(2).

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*. R Foundation for Statistical Computing. http://www.R-project.org/

Salesses, P., Schechtner, K., & Hidalgo, C. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PloS One*, *8*(7).

Solymosi, R., & Bowers, K. (2018). The role of innovative data collection methods in advancing criminological understanding. In G. Bruinsma & S. Johnson (Eds.), *The oxford handbook of environmental criminology* (pp. 210–237). Oxford University Press.

Solymosi, R., Bowers, K., & Fujiyama, T. (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., Bowers, K., & Fujiyama, T. (2017). Crowdsourcing subjective perceptions of neighbourhood disorder: Interpreting bias in open data. *British Journal of Criminology*, *58*(4).

Solymosi, R., Buil-Gil, D., Vozmediano, L., & Guedes, I. (2020). Towards a place-based measure of fear of crime: A systematic review of app-based and crowdsourcing approaches. *Environment & Behavior*.

Spinu, V., Grolemund, G., & Wickham, H. (2020). *Lubridate: Make dealing with dates a little easier*. https://cran.r-project.org/web/packages/lubridate/index.html

Tester, G., Ruel, E., Anderson, A., Reitzes, D., & Oakley, D. (2011). Sense of place among atlanta public housing residents. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, *88*(3).

Traunmueller, M., Marshall, P., & Capra, L. (2015). Crowdsourcing safety perceptions of people: Opportunities and limitations. In T. Liu, C. Scollon, & W. Zhu (Eds.), *SocInfo2015: Social informatics* (pp. 120–135). Springer.

Warr, M. (2000). Fear of crime in the united states: Avenues for research and policy. *Criminal Justice*, *4*(4).

Welsh, B., & Farrington, D. (2004). Surveillance for crime prevention in public space: Results and policy choices in britain and america. *Criminology & Public Policy*, *3*(3).

Wickham, H., Chang, W., Henry, L., Pedersen, T., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2020). *Ggplot2: Create elegant data visualisations using the grammar of graphics*. https://CRAN.R-project.org/package=ggplot2

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

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

Williams, M., Burnap, P., & Sloan, L. (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., & Kleiber, C. (2015). *Ineq: Measuring inequality, concentration, and poverty*. https://CRAN.R-project.org/package=ineq