Small Area Estimation for Crime Analysis

David Buil-Gil
Department of Criminology, University of Manchester, UK

08/06/2020

Abstract

Victimization surveys provide key information about crimes known and unknown to the police, and are the main source of data to analyze perceived safety and trust in the police. These surveys, however, are only designed to allow the aggregation of responses and production of reliable direct estimates (i.e., weighted means or totals) at very large spatial scales, such as countries or states. Sample sizes are generally too small to produce direct estimates of adequate precision at the increasingly refined spatial scales of the criminology of place. Model-based small area estimation may be used to increase the reliability of small area estimates produced from victimization surveys. Small area estimation techniques are designed to produce reliable estimates of parameters of interest (and their associated measures of error) for areas for which only small or zero sample sizes are available. In 2008, the US Panel to Review the Programs of the Bureau of Justice Statistics recommended the use of small area estimation to produce subnational estimates of crime. Since then, these techniques have been applied to study many variables of interest in criminology. This chapter introduces theory and a step-by-step exemplar study in `R` to show the utility of small area estimation to analyze crime and place. Small area estimates of trust in the police are produced from European Social Survey data.

Keywords: Confidence in policing, European Social Survey, crime mapping, open data, GIS

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 ID: David Buil-Gil: 0000-0002-7549-6317.

1. Introduction

The study of crime and place is moving toward analyzing smaller levels of geography than ever before. Micro-level maps of crime are used to foreground those places where crime concentrates, and to design spatially-targeted strategies to tackle crime and disorder (Braga et al., 2018; Groff et al., 2010; Weisburd et al., 2012). The move toward micro-level crime mapping has also provoked those researchers preoccupied with the study of emotions about crime and perceptions about the police, shifting their attention toward small levels of analysis to study the effect of the immediate environment and social context on people's perceptions and emotions about crime and the police (e.g., Solymosi et al., 2015; Williams et al., 2019). These issues have become topics of major interest for researchers, crime analysts and police administrations, who require precise - and reliable - maps of their spatial distribution to develop more accurate theoretical explanations and more efficient evidence-based policing strategies.

Police-recorded crimes and calls for police services are the main sources of data used to identify those micro places where crime is more prevalent. Crimes known to police, however, may be affected by missing data due to underreporting affecting some areas more than others (Brantingham, 2018; Xie & Baumer, 2019). Victimization surveys provide key information to account for crimes known and unknown to the police, and are the main source of information to analyze emotions about crime and perceptions about the police (Rosenbaum & Lavrakas, 1995; Xie & Baumer, 2019). The two main victimization surveys are probably the

National Crime Victimization Survey (NCVS) in the United States, and the Crime Survey for England and Wales (CSEW) in the UK. However, these surveys are not designed to allow researchers and practitioners to aggregate responses at the level of small geographies. Sample sizes are large enough to produce direct outputs only for large spatial scales, such as countries or states, but small levels of geography suffer from small sample sizes. In more precise terms, surveys are generally designed to allow producing reliable direct estimates (i.e., weighted means or totals) for large areas included in the sampling design, but directly aggregating data from a few respondents in each small area will inevitably lead to imprecise and unreliable maps. Model-based small area estimation (SAE) techniques may be applied to increase the reliability of small area estimates of parameters of criminological interest produced from survey data (Rao & Molina, 2015).

This chapter introduces theory and a step-by-step exemplar study in `R` (R Core Team, 2020) to show the utility of SAE to analyze crime and place. In the exemplar study we illustrate how to produce small area estimates of trust in the police from European Social Survey (ESS) data.

2. Foundations of small area estimation

Pfeffermann (2013) defines SAE as a group of techniques designed to "produce reliable estimates of characteristics of interest such as means, counts, quantiles, etcetera, for areas or domains for which only small samples or no samples are available" (p. 40). SAE is an umbrella term used to classify many different techniques developed to deal with different types of data, but all aimed to improve the reliability of estimates produced for areas with small sample sizes (see a review of the main SAE techniques in Rao & Molina, 2015). In this regard, those who use SAE techniques use the term *'small areas'* to describe not only low-level geographies, but any area whose sample size is too small to allow the production of direct estimates with adequate precision. In the exemplar study presented below, for example, we will produce small area estimates of trust in the police in European regions. These are not small units of analysis, but suffer from small sample sizes in many areas.

2.1 Direct estimation

As described above, direct estimators use only survey data recorded in each area (including survey responses to the variable of interest and survey weights) to obtain weighted means or totals in each area. Survey weights are calculated and included in most surveys to adjust selected samples to the target population, thus making survey data more representative. We will call *'direct estimates'* those estimates computed only from survey data. Direct estimates will be unreliable in all those areas where only small samples were recorded by the original survey. We will use the Coefficient of Variation (CV) of the direct estimate in each area, which is the ratio between the standard deviation of the weighted mean and the weighted mean, to define the *reliability* of each direct estimate. Direct estimates may not be reliable enough in many areas, and thus we will need to apply *indirect* SAE techniques to calculate more reliable estimates.

2.2 Model-based estimation

Indirect SAE is also known as *'model-based'* SAE, since these techniques make use of explicit linking models to 'borrow strength' across related areas. In essence, these techniques link information recorded by the original survey in each area to auxiliary variables (generally) registered by administrative agencies or census data, and use this information to increase the reliability of our estimates in all areas. For example, if we know that the chances of falling victim to crime are associated with the level of neighborhood deprivation, we can estimate a model that links the crime victimization recorded in the survey to known poverty measures in each place, thus producing higher estimates of crime victimization in deprived areas, and lower estimates in wealthier areas. These auxiliary variables are generally called *covariates*. The quality of the covariates used and the accuracy of the linking model will be key to obtain reliable small area estimates.

When we use explicit models to link information available for all sampled units with unit-specific covariates, then we are using a *'unit-level'* SAE approaches. On the other hand, if our linking model is estimated by relating the area-level direct estimates of our variable of interest to area-level covariates, then we would be using *'area-level'* SAE. Area-level SAE approaches are generally preferred when our variable of interest is particularly affected by the contextual characteristics of each area (Namazi-Rad & Steel, 2015). The study of crime and place has shown that crime-related variables are strongly associated with the characteristics of small areas (e.g., Brunton-Smith & Jackson, 2012; Weisburd et al., 2012), and thus here we focus on the use of area-level SAE.

2.2.1 Regression-synthetic estimation

More specifically, when some sort of regression model is used to link survey data to covariates in order to estimate regression coefficients (and sometimes other model parameters) and compute area-level predictions from these, then we are producing regression-based synthetic estimates. These can be based on linear, logistic, multilevel or more complex modeling approaches. Synthetic estimates can be obtained for all areas (even those with zero sample sizes), but these are too much model-dependent and susceptible to model misspecification (Levy, 1979; Rao & Molina, 2015). Synthetic estimates suffer from a high risk of bias. We will illustrate how to obtain area-level regression-based synthetic estimates later on, and we will also show why these may be unreliable due to their high risk of bias.

2.2.2 The Empirical Best Linear Unbiased Predictor (EBLUP)

A composite estimator that combines the direct and synthetic estimate in each area is the preferred approach to rectify all these deficiencies. The basic SAE approaches consist precisely of optimal combinations between two components given by the direct estimate, which will be more reliable in some areas than others, and a synthetic estimate. Since sample sizes are always larger in some areas than others, calculating direct estimates from survey data will generate more reliable estimates in areas with larger samples than in areas with small samples. These composite estimators attach more weight to the direct estimate when its sampling variance is small, while more weight is attached to the synthetic estimator when the variance of the direct estimate is larger. This is done in each area, obtaining an optimal combination between the direct and synthetic estimate depending on the level of reliability of the direct estimate. One of the most widespread area-level composite estimators in SAE is the EBLUP, which was developed originally by Fay & Herriot (1979). In this chapter we will produce EBLUP estimates of trust in the police in Europe.

2.3 Calculating the uncertainty of estimates

One of the main advantages of using SAE is that a great deal of work has been carried out on how to estimate the measure of uncertainty of each small area estimate produced in each area. These measures of uncertainty are usually presented as Mean Squared Errors (MSE) or Relative Root Mean Squared Errors (RRMSE). The MSE is the averaged squared error of an estimate, representing the difference between the estimate value and the expected true value of what is measured. MSE is computed to account for both the variance of estimates (i.e., spread of estimates from one sample to another) and their bias (i.e., distance between the estimated value and the true value). The MSE is always non-negative, and values closer to zero indicate a higher level of reliability of our estimate. We obtain the RRMSE by taking the square root of the MSE (i.e., to calculate the Root Mean Squared Error, RMSE), and then dividing it by the corresponding estimate. RRMSEs are generally presented as percentages.

Calculating the RRMSE of estimates allows examining which SAE method produces the most reliable estimates, but it also allows identifying which specific estimate in which specific area may suffer from inadequate reliability. Different statistical agencies use different threshold points to decide whether an estimate is reliable enough. Here we consider that estimates with a RRMSE smaller than 25% are reliable, estimates with a RRMSE between 25% and 50% can be used with caution, and estimates with a RRMSE larger than 50% are too unreliable to be used (Commonwealth Department of Social Services, 2015).

We will use the CV of direct estimates as the measure of uncertainty, since it corresponds to the RRMSE of model-based estimators (Rao & Molina, 2015). To compute the RRMSE of our EBLUP estimates we follow the analytical procedure described in Datta & Lahiri (2000).

3. Small area estimation applications for crime analysis

SAE may be of great value for the study of crime and place: to estimate the geographical distribution of crimes known and unknown to police, and to produce detailed maps of crime-related perceptions and emotions. This is the reason why, in 2008, the US Panel to Review the Programs of the Bureau of Justice Statistics (BJS) recommended the use of model-based SAE to produce subnational estimates of crime rates: "BJS should investigate the use of modelling NCVS data to construct and disseminate subnational estimates of crime and victimization rates" (Groves & Cork, 2008, p. 8). This work was started by Robert E. Fay and colleagues to produce estimates of victimization rates for states and large counties in the US (Fay & Diallo, 2015, 2012). The need for applying SAE to estimate crime has also been acknowledged by the Australian Bureau of Statistics (Tanton et al., 2001) and Statistics Netherlands (Buelens & Benschop, 2009).

Some researchers have applied different regression-based synthetic estimators to produce small area estimates of crime and disorder, but, as discussed above, these are known to suffer from a high risk of producing biased estimates (Levy, 1979; Rao & Molina, 2015).

Others have used the basic unit-level or area-level EBLUP, or temporal and spatial extensions of the area-level EBLUP, to produce estimates of crime rates. Buelens & Benschop (2009) applied the area-level EBLUP to produce estimates of victimisation rates in police zones in Netherlands. Fay and colleagues developed an area-level dynamic SAE approach and produced estimates of crime rates in states and large counties in the US (Fay & Diallo, 2015, 2012). D'Alo et al. (2012) made use of the unit-level and area-level EBLUP to produce estimates of violence against women at a regional level in Italy. And Buil-Gil, Moretti, et al. (2019) and Buil-Gil, Medina, et al. (2019) applied a spatial extension of the area-level EBLUP to produce estimates of worry about crime and perceived neighborhood disorder, respectively. Here we will show how to produce area-level EBLUP estimates of trust in the police in Europe.

4. Small area estimation of trust in the police: Step-by-step example in `R`

In this section we illustrate how to produce small area estimates with real-world data. More specifically, we will use data recorded by the ESS to produce estimates of trust in the police across regions in Europe. In practice, crime and place researchers will be more interested in applying SAE to estimate crime victimization, perceived safety and trust in the police at smaller spatial scales, but survey data with such level of spatial granularity are usually subject to great levels of scrutiny, and survey administrators tend not to publish micro-data at the level of small geographies. Access to such data is generally subject to special permissions and cannot be shared openly. The reproducible example shown here will illustrate how to apply SAE methods using open data, so it can be followed and used in a variety of other datasets and variables of interest.

The exemplar study is designed for you to acquire a number of different practical skills, which include:

- Downloading and exploring survey data.
- Analyzing issues around survey data coverage of small areas and small sample sizes.
- Producing direct estimates for small areas and exploring their reliability.
- Producing area-level EBLUP estimates, which involves several steps:
  - accessing reliable area-level covariates,
  - estimating area-level regression models and obtaining synthetic estimates,
  - calculating EBLUP estimates and their RRMSE,

- visualizing the spatial distribution of EBLUP estimates,
- visualizing the improved reliability of EBLUP estimates over direct estimates, and
- producing model diagnostics to check if our model-based estimates are unbiased.

4.1 European Social Survey

In this exemplar study we will use ESS data. The ESS is a biannual cross-national survey designed to measure social attitudes, beliefs and behaviors in Europe. It has been conducted since 2001 in more than 35 countries, and is designed to be representative of all individual residents aged 15 or older who live in private households in each participant country, regardless of their nationality, citizenship or language. The ESS allows for cross-national and cross-sectional comparisons of crime-related issues such as the trust in police services, worry about crime and crime victimization experienced in the last 5 years. For instance, it includes the following variables of interest for crime analysts:

1. *"Have you or a member of your household been the victim of a burglary or assault in the last 5 years?"*
2. *"How safe do you – or would you – feel walking alone in this area after dark?"*
3. *"Using this card, please tell me on a score of 0-10 how much you personally trust each of the institutions I read out [...]"*: *"[...] the legal system"* and *"[...] the police"*.

These measures have previously been used to study victimization, perceived safety, trust in the police, and trust in the legal system (e.g., Hough et al., 2014; Hummelsheim et al., 2011; Kääriäinen, 2007), but there are many other questions that may also be of interest (e.g., racism, homophobia). The whole ESS questionnaire is available here: https://www.europeansocialsurvey.org/docs/round8/fieldwork/source/ESS8_source_questionnaires.pdf.

With regards to the sampling design, participant countries are responsible for producing their own national sampling designs following some common sampling principles. Namely, respondents must be selected following strict random probability techniques at every stage, sampling frames can be individuals, households or addresses, quota sampling is not allowed, and non-responding units cannot be replaced. Moreover, every country must select at least 1,500 effective respondents (or at least 800 in countries with less than 2 million citizens). As a consequence, countries with very different population sizes may select similar sample sizes, and geographical levels below countries (e.g., regions, counties) are not planned by the original sampling design in most countries and suffer from small sample sizes.

4.1.1 Download European Social Survey data

ESS data can be downloaded from their website (https://www.europeansocialsurvey.org/). But we can also download ESS data directly into our `R` system using the `essurvey` package developed by Cimentada (2019). This package is designed to facilitate loading ESS survey data into `R`. It allows users to select the countries and years they are interested to analyze and load survey data directly in `R`. If this is the first time we are using this package, we need to install it first by using the `install.packages()` function.

```r
install.packages("essurvey") #install essurvey package
```

Once it is installed, we can load the package into our `R` environment using the `library()` function.

```r
library(essurvey) #load essurvey package
```

In order to access ESS data in `R`, first we need to create our own personal account in the ESS online portal. ESS users need to register only once, and after that they can have open access to ESS data as many times as they wish. In order to sign up for a ESS account, we need to access the ESS website and create a new account with our personal details: https://www.europeansocialsurvey.org/user/new. Filling the online registration form takes less than two minutes, and once it is completed we will receive an email to confirm our registration.

Once we are registered in the ESS platform, we can directly import all ESS data into R. In this exercise we will download and analyze data from the 8th edition of ESS, which was published in 2016. We use the function set_email() from essurvey to save our email (the email account registered in the ESS platform) as a new environment variable, and then run the import_rounds() function to load ESS data from all participant countries. This may take a few seconds.

```
set_email("your_email@domain.com") #change by your email

ess <- import_rounds(rounds = 8)   #load ESS data
```

ESS data from 44387 respondents across 20 different countries are now loaded into R, and we can begin exploring and analyzing these survey data. If we want to see how the dataset looks like, we can use the View() function.

4.2 Descriptive analyses

In this exemplar study we will analyze ESS data about trust in police services, following previous research conducted by Kääriäinen (2007), Staubli (2017) and others. The variable name is trstplc, which uses a Likert scale variable from 0 to 10, where 0 indicates the lowest level of trust, and 10 is the maximum value. We will begin checking how this measure of trust in the police looks like. We will use the summary() function to obtain the summary statistics of this variable. First, however, we need to transform the class of this variable from 'haven_labelled' (i.e., SPSS variable handled in R) to numeric, to facilitate handling the data in R. We transform the variable using the mutate() function from dplyr package (Wickham, François, et al., 2020) and convert it into numeric using the as_numeric() function.

```
library(dplyr) #load dplyr package

ess <- ess %>%                             #set dataset to use
  mutate(trstplc = as.numeric(trstplc)) #transform column to numeric

summary(ess$trstplc) #print summary statistics
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   5.000   7.000   6.399   8.000  10.000     320
```

We can see that the average score of trust in police in Europe is 6.4, and the median value is 7. We can use the same summary() function to analyze the values of trust in other social institutions, such as the legal system (variable trstlgl), politicians (trtplt), political parties (trtprt), the country's parliament (trstprl) or the United Nations (trstun). On average, Europeans appear to have more trust in the police than in other key political and legal institutions. We can also see that 320 persons did not answer this question (i.e., 'NAs').

We can obtain some more detailed information about the citizens' trust in the police by counting the frequency of respondents that chose each score, and creating a bar plot to visualize their distribution. We will use functions from the the packages dplyr and ggplot2 (Wickham, Chang, et al., 2020) for this. More specifically, we use the group_by() function from dplyr to create groups of respondents based on their score of trust in police, and use the functions summarize() and mutate() from the same package to save the results in two columns showing the number (n) and proportion (prop) of respondents in each category. We save this table into a new data frame object called trust_poli.

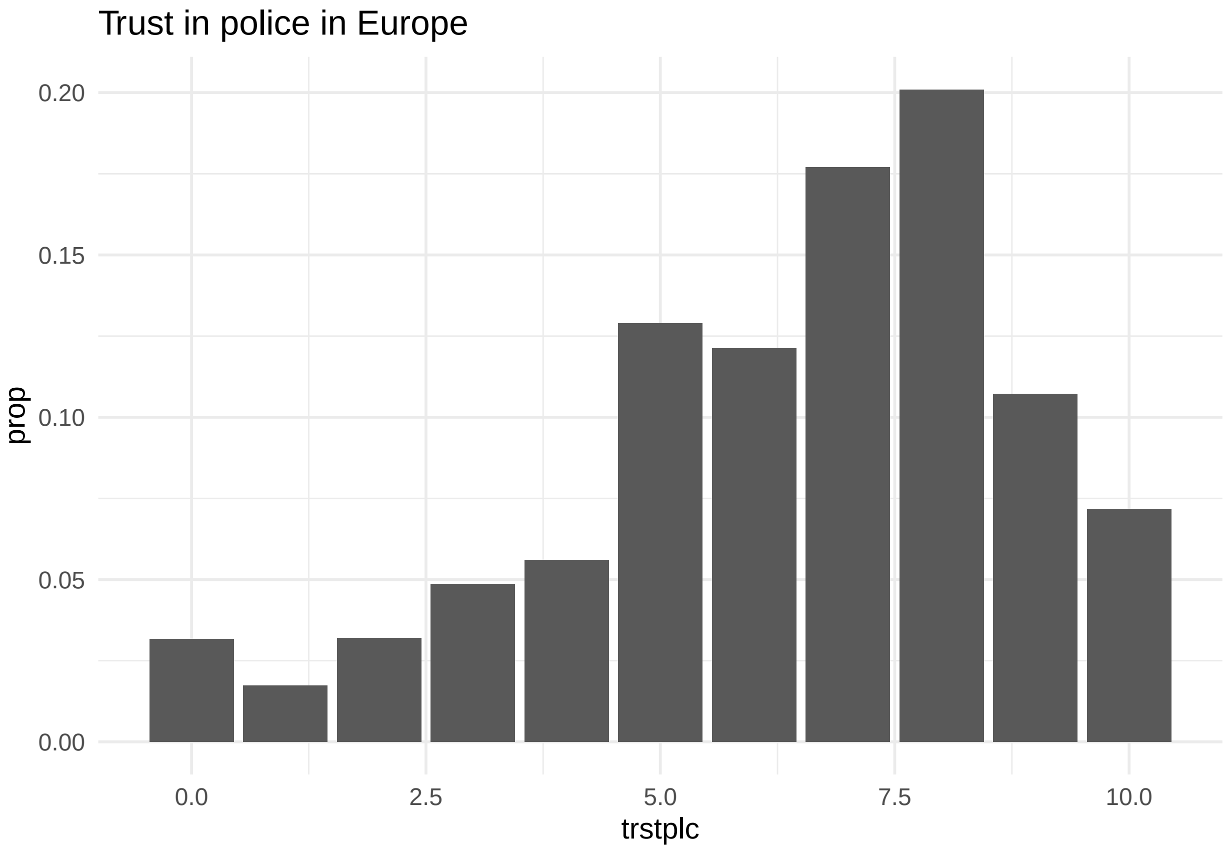
```
trust_poli <- ess %>%       #set dataset to use
  group_by(trstplc) %>%     #group by score of trust
  summarize(n = n()) %>%    #number of respondents per group
  mutate(prop = n / sum(n))#proportion respondents per group
```

Then, we use the `ggplot()` and `geom_bar()` functions from `ggplot2` to create a bar graph of the number of responses per score of trust. Before plotting this visualization, we will run the function `theme_set(theme_minimal())` to set a basic, neat theme for all our plots.

```
library(ggplot2) #load ggplot2 package

theme_set(theme_minimal()) #set white theme for plots

ggplot(data = trust_poli,               #set dataset to use
       aes(x = trstplc, y = prop)) +   #set variables for x and y axis
  geom_bar(stat="identity") +          #plot bars showing values in data
  ggtitle("Trust in police in Europe")#change title
```



We see that only a small proportion of respondents have a low trust in the police, whereas most appear to trust the police quite a lot. This plot, nevertheless, is likely to hide internal heterogeneity between European regions and countries. Based on this bar graph alone, we do not have enough information to be able to know if residents in all participant countries have similar levels of trust in the police, or whether respondents with very low or very high trust in the police concentrate in some regions but not others.

4.2.1 Recode the measure of trust in police work

In this exemplar study, we are particularly interested in analyzing which regions in Europe present especially low or high levels of trust in the police in comparison to other European countries. One way to do this is by estimating the proportion of residents in each region who have a level of trust above the average in Europe. Our final estimates will show a value between 0 and 1 representing the proportion of residents that have more trust in the police than the European average. For instance, a value of 0.6 in a given region would indicate that 60 percent of its residents have more trust in the police than the average of all European citizens.

Thus, we need to recode our variable of interest. We will use the `mutate()` and `if_else()` functions from `dplyr` to do this. Those respondents with a score above or equal to the European mean will be given a value `1`, whereas the others will be assigned a value of `0`. This will facilitate the interpretation of our results, but future research can explore producing estimates from the original 0-to-10 Likert scale. We will also delete all those respondents who did not answer this question (i.e., *NA*s).

```
ess <- ess %>%           #set dataset to use
         #if trust above or equal to mean, assign 1, if not 0
  mutate(trstplc = if_else(trstplc >= mean(trstplc, na.rm = T), 1, 0)) %>%
  filter(!is.na(trstplc))#delete NAs
```

We can also use the `group_by()` and `summarize()` functions seen above to explore our recoded variable. We can see, for example, that 24721 out of 44067 respondents (i.e., 56.1% of participants) have more trust in the police than the average in Europe.

4.3 Exploring spatial data: Coverage and sample sizes

As mentioned above, the ESS sample is designed to allow the production of reliable direct estimates at the country level, but samples recorded at smaller scales (e.g., regions, cities) may be too small to allow producing direct estimates with adequate precision in all areas. We can check how large ESS sample sizes are in each region (variable `region`) using the functions `filter()`, `group_by()` and `summarize()` from `dplyr` to create a summary table in a new data frame that we will call `sample_region`. We will also remove one unit that has missing data in the `region` column (i.e., value 99999). Then, we can use the `summary()` function to print the summary statistics of area sample sizes.

```
sample_region <- ess %>%     #set dataset to use
  filter(region != 99999) %>%#filter out NAs
  group_by(region) %>%       #categories based on region
  summarize(n = n())         #calculate sample size

summary(sample_region$n) #print summary statistics
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     7.0    60.0   115.0   160.8   209.8  2524.0
```

The average sample size per region is 160.8, which is quite large but may be insufficient to produce reliable direct estimates in all areas. We can also see that there are areas with very small sample sizes (the minimum area sample size is 7), where we cannot simply rely on direct estimation techniques to obtain small area estimates of adequate precision.

4.3.1 Checking the spatial scale reported in each country

We also need to consider that participant countries can decide whether they want to publish their recorded data at the level of NUTS-1, NUTS-2, NUTS-3 or smaller spatial scales. NUTS is the acronym of *'Nomenclature of Territorial Units for Statistics'*, and it refers to spatial scales used by the European Union and

Eurostat (the statistical office of the European Union) for policy making and statistical reporting purposes. NUTS are basically a way to organize European countries in regions and subregions. In England, for example, NUTS-1 are statistical regions, NUTS-2 are counties (and groups of districts in London), and NUTS-3 are generally unitary authorities (some grouped). Whereas some participant countries publish their data at the level of NUTS-2, others decide to report information for NUTS-1 or NUTS-3 areas. We can check which level of aggregation is published by each participant country in the ESS website: https://www.europeansocialsurvey.org/data/multilevel/guide/essreg.html. We can also run the following lines of code and print this information directly in `R`.

```
ess %>%                       #set dataset to use
  group_by(regunit, cntry) %>%#group by spatial scale and country
  summarize(n = n())          #print sample size per country
```

We see that many countries publish data at the NUTS-2 level, but others participant countries publish their micro-data for NUTS-1 and NUTS-3 areas. We will aggregate all data at the NUTS-2 level and produce estimates of trust in the police at this spatial scale (with the exception of Germany and the UK, who only publish data for NUTS-1).

4.3.2 Converting spatial data into NUTS-2

In order to convert the spatial information provided by ESS into NUTS-2 geographies, we first need to load a lookup table that details which NUTS-3 areas are part of which NUTS-2 geographies. I have previously created and saved a lookup table in csv format in an open access Github repository, but similar tables are also available in other formats from the ESS platform: https://www.europeansocialsurvey.org/data/multilevel/guide/bulk.html. In order to load the lookup table in `R`, we can use the `getURL()` function from `RCurl` (CRAN team, 2020) and `read.csv()`.

```
library(RCurl) #load RCurl package

#save URL address as character value
url_lookup <- getURL("https://raw.githubusercontent.com/davidbuilgil/SAE_chapter/master/data/NUTS_lookup

lookup <- read.csv(text = url_lookup) #load lookup table
```

Now, we can create a new column in the original ESS data that specifies the regions for which we aim to produce small area estimates of trust in the police. To do this, first, we merge the lookup table with the original ESS data using a `left_join()` function, and then we create a new column called `domain` which shows the NUTS-2 areas (or NUTS-1 in Germany and UK) for which we will produce estimates.

```
ess <- ess%>%                                #set dataset to use
  left_join(lookup,
            by = c("region" = "nuts3")) %>%  #merge lookup into ESS dataset
  rename(domain = nuts2) %>%                 #rename NUTS2 variable as domain
  mutate(domain = as.character(domain),      #convert NUTS2 into character
         domain = if_else(is.na(domain),
                          region, domain))%>%#copy NUTS1 if there is no NUTS2
  filter(!(domain == 99999))                 #delete NAs
```

Now our data is clean and ready to be used to produce estimates of trust in the police at a regional level.

4.4 Producing direct estimates

We can begin by producing direct estimates in each area. As discussed before, direct estimators make use of original survey data and survey weights to obtain design-unbiased estimates in each small area, but direct estimates may be too unreliabile in those areas with small sample sizes. We will produce direct estimates of trust in the police for European regions, but it is very likely than many estimates will not show adequate levels of precision. Model-based SAE approaches are needed when direct estimates are not precise enough.

In order to produce small area estimates, we will use the `sae` package (Molina & Marhuenda, 2020). We need to install it and load it into our `R` system.

```
library(sae) #load sae package
```

Direct estimators take into account the population size in each area, and assume that survey weights adjust our sample size to the total population in each area. In other words, we need to know the number of residents in each region, and ensure that our survey weights adjust the area sample size to the population size.

4.4.1 Download data about population sizes

I have previously downloaded the regional population sizes from Eurostat and uploaded a clean dataset onto a Github repository. Downloading data from sources of official statistics, such as Eurostat, usually means having to spend some time cleaning and wrangling the data. For the purpose of this exercise, I have downloaded, cleaned and saved the data previously, but later we will also see how to load Eurostat data into our `R` environments.

```
#save URL address as character value
url_pop <- getURL("https://raw.githubusercontent.com/davidbuilgil/SAE_chapter/master/data/popsize.csv")

pop <- read.csv(text = url_pop) #load population size
```

We have almost all the information necessary to produce our direct estimates: *(a)* the variable of interest (column `trstplc` in `ess` dataset), *(b)* the area population size (`pop2016` in `pop` dataset), and *(c)* spatial information that matches both datasets. Nevertheless, as introduced above, direct estimators also require the use of survey weights that adjust our sample size to the population size in each.

4.4.2 Readjust ESS weights

The survey weights published by ESS are not designed to let respondents represent a specific number of citizens in the population, but instead they were computed to adjust chances of selection of every unit in the sample to the population characteristics. You can find a detailed guide about ESS survey weights here: https://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf. We will readjust our survey weights so that weighted respondents coincide to known population totals, while maintaining the design of the original weights that adjust for sample selection bias and unequal population sizes across countries. We can do this by running the following lines of code:

```
ess_w_area <- ess %>%                      #set dataset to use
  filter(domain %in% pop$domain) %>%       #filter areas from pop dataset
  group_by(domain) %>%                     #create groups by region
  summarise(w_sum = sum(pspwght * pweight))#sum weights per region

ess <- ess %>%                               #set dataset to use
  filter(domain %in% pop$domain) %>%         #filter areas from pop dataset
  left_join(ess_w_area, by = "domain") %>%   #merge sum of weights with ESS
  left_join(pop, by = "domain") %>%          #merge region population sizes
  mutate(weight = pspwght * pweight,         #compute cross-national weights
         weight = (weight * pop2016) / w_sum)#adjust weights to pop size
```

4.4.3 Calculate direct estimates

After a few steps, we finally have all necessary information to produce our direct estimates of trust in the police. We use the `direct()` function from `sae` to produce direct estimates in each region. It will also produce the CV of each estimate, which will be used to assess the reliability of these direct estimates.

```
dir <- direct(y       = ess$trstplc,#set variable of interest
              dom     = ess$area,   #areas to produce estimates
              sweight = ess$weight, #survey weights
              domsize = pop[,2:3])  #population size
```

4.4.4 Exploring direct estimates

Once we have produced our direct estimates of trust in the police, we can see how these look like by using the `View()` function. We can also obtain some summary statistics of our direct estimates using the `summary()` function introduced before.

```
summary(dir$Direct) #summary statistics of direct estimates
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1783  0.4877  0.5815  0.5838  0.6704  0.9006
```

Our direct estimates show that, across all areas, the average percentage of residents who have more trust in the police than the European average is 58%, but there is a large variation across areas. For instance, the region with the smallest direct estimate indicates that only 18% of residents have more trust in the police than the average in Europe, whereas there is one region where 90% of residents have more trust in the police than the European average.

However, we do not know yet how reliable these direct estimates are. We can explore the direct estimates' CV by opening this data frame (`View()` function), but we can also print some summary statistics that will give us a hint of the level of reliability of our direct estimates.

```
summary(dir$CV) #summary statistics of CV
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.843   8.996  11.975  14.095  16.617  50.194
```

As you can see, the average CV of our direct estimates (14.09%) is quite small according to SAE standards, but the maximum value of CV is 50.19, which indicates that the direct estimate in at least one area is highly unreliable. If we observe our direct estimates, we will see that 60 direct estimates have measures of CV larger than 15%, 17 of them have a CV larger than 25%, and 1 of them has a CV larger than 50%. These results are not the end of the world, but we need to do our best to improve the reliability of small area estimates in those areas where the CV of direct estimates show inadequate levels of reliability. We need model-based SAE.

For now, we will merge our direct estimates into the dataset of area-level information by using the `left_join()` function from `dplyr`.

```
pop <- pop %>%                             #set dataset to use
  left_join(dir, by = c("area" = "Domain"))#merge direct estimates
```

4.5 Downloading area-level covariates

In order to estimate area-level models and produce model-based small area estimates of trust in the police, we need area-level covariates associated with our variable of interest. We may need to conduct a preliminary literature review to explore which contextual features have been used to explain our variable of interest. For example, in our case, we can read key articles and books analyzing the contextual predictors of trust in the police in European countries, such as Jackson et al. (2013), Kääriäinen (2007) and Staubli (2017), but also in other contexts (e.g., Cao et al., 2012). For instance, we can see that researchers have used measures related to crime rates, education level, income and level of democracy, among others, to explain the meso- and macro-level spatial distribution of trust in the police. We will download and use data about some of these covariates to estimate our area-level models and produce model-based estimates of trust in the police.

We can download various area-level covariates from Eurostat using the `eurostat` package (Lahti et al., 2020). Eurostat publishes large datasets of social, economic and demographic information for European countries and regions. We can use the `search_eurostat()` function to search predefined key words associated with variables of interest for our study, such as *"education"* or *"offender"*. This function will return a list of all available datasets including our keywords. We can save this list and explore which datasets may be more suitable for our area-level models. For example, we may want to know if education levels and crime rates are somehow associated with the regional levels of trust in the police. I have done this search and found various variables of interest, but you can also try this at home, and probably you will find other interesting datasets to be used in this study.

```
library(eurostat) #load eurostat package

eurostat_edu <- search_eurostat("education") #search data about education
eurostat_cri <- search_eurostat("offender")  #search data about crime
```

We can see, for example, that our search has found 1006 datasets that contain the word *"education"*, and 4 that contain the word *"offender"*. We will need to spend some time searching which covariates are the most suitable for our study. Once we know the codes of the datasets we are interested to use, we can use the `get_eurostat()` function to import these into our `R` system. For example, the dataset `edat_lfs_9918` includes information about the proportion of citizens between 15 and 64 that have a higher education degree in each region. We can download this dataset and see what it looks like:

```
he <- get_eurostat(id = "edat_lfs_9918") #download data from Eurostat
```

This is just one possible source of data that we can use to find covariates for our area-level models. The ESS website also publishes interesting covariates at the different spatial scales (https://www.europeansocialsurvey.org/data/multilevel/guide/bulk.html). In any case, we will need to spend some time wrangling and subsetting these sources data to make sure we can use them as covariates in our models.

For the purpose of this exemplar study, I have previously searched for datasets of interest, downloaded and cleaned their data, and merged all covariates into a unique dataset. I have also used multiple imputation via bootstrapping and predictive mean matching to impute a few missing values in some regions (see original codes used to impute missing values here: https://github.com/davidbuilgil/SAE_chapter/blob/master/other_codes/codes_imputation.R). We can load the dataset with our clean and ready-to-use covariates into `R` using the functions provided by `RCurl` package, but you can also spend some time trying to find better, more suitable covariates using the `eurostat` package.

```
#save URL address as character value
url_covs <- getURL("https://raw.githubusercontent.com/davidbuilgil/SAE_chapter/master/data/covs_short_i

covs <- read.csv(text = url_covs) #save covariates

pop <- pop %>%                   #set dataset to use
  left_join(covs, by = "domain") #merge covariates with our data
```

We can open this file using the `View()` function and see that it includes six area-level covariates that we will use in our models:

- *'fem_p_16'*: Proportion of females residing in each area in 2016.
- *'gdp_eurhab_16'*: Gross Domestic Product (GDP) per capita in 2016, in euros (€).
- *'robb_r_10'*: Robbery rate by 10,000 population in 2010.
- *'burg_r_10'*: Burglary rate by 10,000 population in 2010.
- *'he_p_16'*: Proportion of population with a higher education degree in 2016.
- *'medage_16'*: Median age in 2016.

One may notice that all these covariates represent data from 2016, with the exception of our two covariates measuring crime rates (from 2010). 2010 was the last year in which Eurostat published crime statistics for European regions. Since we are particularly interested in knowing the effect of crime rates on the trust in the police, we accept these covariates as proxy measures of crime levels for the purpose of this study. However, we should be using covariates from the same year as our survey data when possible.

4.6 Fitting area-level models and obtaining synthetic estimates

All our area-level variables (i.e., outcome measures measured by the direct estimates, and covariates) are now clean and ready to be used to produce small area estimates. The first step, as discussed before, is fitting an area-level model that links our direct estimates with our covariates. We will use the `lm()` function to fit our area-level linear model, and we can see the results of our model using the `summary()` function.

```
#estimate area-level linear model
model <- lm(Direct ~ fem_p_16  + gdp_eurhab_16 + robb_r_10 +
                     burg_r_10 +  medage_16    + he_p_16,
            data = pop)
```

As you may have already noticed, we are using covariates with very different scales and dimensions. For example, the covariate about GDP per capita has a minimum value of 7300 and its maximum is 75300, whereas the variable about proportion of females ranges between 0.49 and 0.54, and our covariate measuring rate of robberies varies between 0.35 and 72.64. This is not problematic for regression analysis nor SAE, but we may want to standardize the regression coefficients when presenting the model results, in order to be able to compare the relative importance of each covariate in our model (see Gelman, 2008). We can do this by scaling all our variables using the `scale()` function before fitting the linear model. Model results are shown in Table 1. Scaling our variables does not have any effect on our analyses, but allows us to better understand the effect of each covariate in our regression model. For the purpose of this study, we will consider that any p-value of less than 0.05 is significant.

Table 1: Table 1: Area-level model of trust in the police (standardized coefficients)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.00	0.06	0.00	1.00
Proportion females	-0.26	0.07	-3.57	0.00
GDP per person (€)	0.28	0.08	3.27	0.00
Robbery rate	-0.06	0.07	-0.81	0.42
Burglary rate	-0.02	0.07	-0.29	0.77
Median age	0.16	0.07	2.35	0.02
Proportion HE	0.25	0.08	3.10	0.00

We can see in Table 1 that those European regions with a larger GDP per capita are also those with more trust in the police. We also see that areas with a larger proportion of females show lower values of trust in the police, areas with more citizens with a higher education qualification generally have more trust in the police, and the median age is positively associated with the area-level measure of trust in police (ageing regions trust more the police than areas with younger populations). The two measures of crime rates show a negative coefficient, but these are not statistically significant.

We can also check the model R squared by running `summary(model)$r.squared` and observe that the R Squared equals 0.35, which indicates that our model accounts for 35% of the variation of the direct estimates of trust in police.

The next step is using this area-level linear model to obtain our regression-based synthetic estimates of trust in the police. As discussed above, regression-based estimators can be produced for all areas, but these are not based on a direct measurement of the variable in each area and suffer from a high risk of bias (Levy, 1979). We use the `predict()` function to produce synthetic estimates from our area-level linear model, and then we can merge our synthetic estimates with our main dataset with regional data using the `cbind()` function.

```
synthetic <- predict(model) #predict synthetic estimates

pop <- pop %>%      #set dataset to use
  cbind(synthetic) #merge regression-based estimates
```

4.7 Producing EBLUP estimates

We are now ready to produce our area-level EBLUP estimates of trust in the police. Using the same covariates as before, we can estimate our EBLUP model and produce the EBLUP estimates using the `eblupFH()` function from `sae` package. As discussed above, the EBLUP obtains an optimal combination of direct and regression-based synthetic estimates in each area, giving more weight to the direct estimate dimension when its sampling variance is small, while more weight is attached to the synthetic estimate when the direct estimate's variance is larger. The EBLUP balances for the variance of direct estimates and the risk of bias of regression-based synthetic estimates by producing the optimal combination of these in each area.

When we use the `eblupFH()` function, we need to detail the elements of the model inside the argument part called `formula`, the variance of direct estimates (i.e., the square of the standard deviation, SD) inside the `vardir` argument, and the type of fitting method inside `method`. Here we use a Restricted Maximum Likelihood (REML) fitting method, which takes into account for the loss in degrees of freedom derived from estimating the model coefficients, but other approaches (e.g., Maximum Likelihood, ML, or Fay-Herriot, FH) should produce similar results.

```
eblup <- eblupFH(formula = pop$Direct        ~ pop$fem_p_16  +
                           pop$gdp_eurhab_16 + pop$robb_r_10 +
                           pop$burg_r_10     + pop$medage_16 +
                           pop$he_p_16,
                 vardir  = pop$SD^2,
                 method  = "REML")
```

We can get the summary of the EBLUP model results by running `eblup$fit`, and observe whether the p-values of our EBLUP model have changed in comparison to the p-values of our linear model estimated before. We can also check the summary statistics of our EBLUP estimates of trust in the police.

```
summary(eblup$eblup) #print summary statistics of EBLUPs
```

```
##        V1
##  Min.   :0.2928
##  1st Qu.:0.5074
##  Median :0.5870
##  Mean   :0.5790
##  3rd Qu.:0.6514
##  Max.   :0.8628
```

Our estimates indicate that the European region with the lowest trust in the police is defined by 71% of residents who have a lower trust in the police that the average in Europe, whereas there is another region where 86% of citizens have more trust in the police that the average in Europe. The average level of trust in the police is 58% (sd = 0.11).

Finally, we can merge our EBLUP estimates with our main dataset of area-level information. We use the `cbind()` function to merge these data as a new column.

```
pop <- pop %>%                 #set dataset to use
  cbind(eblup$eblup) %>%       #merge data into main dataset
  rename(eblup = "eblup$eblup")#change name of column
```

4.7.1 Mapping the trust in police work in Europe

Now that we have computed our EBLUP estimates of trust in the police, we can visualize these in a map to analyze their geographic distribution. We will need a shapefile of NUTS-2 regions (NUTS-1 in the UK and Germany), to which we will attach our estimates and print the map of trust in police. I have previously prepared our shapefile of European regions and uploaded it onto a Github repository, so we can just load it into our `R` environment by using the `sf` package (Pebesma, 2020) and functions from the `RCurl` package.

```
library(sf) #load sf package

#save URL address as character value
url_nuts <- getURL("https://raw.githubusercontent.com/davidbuilgil/SAE_chapter/master/shapefile/nuts_es

nuts <- st_read(url_nuts) #load shapefile
```

We will also use the `st_crs()` function from `sf` to change the Coordinate Reference Systems (CRS) to ED79, which shows a much clearer visualization of Europe.

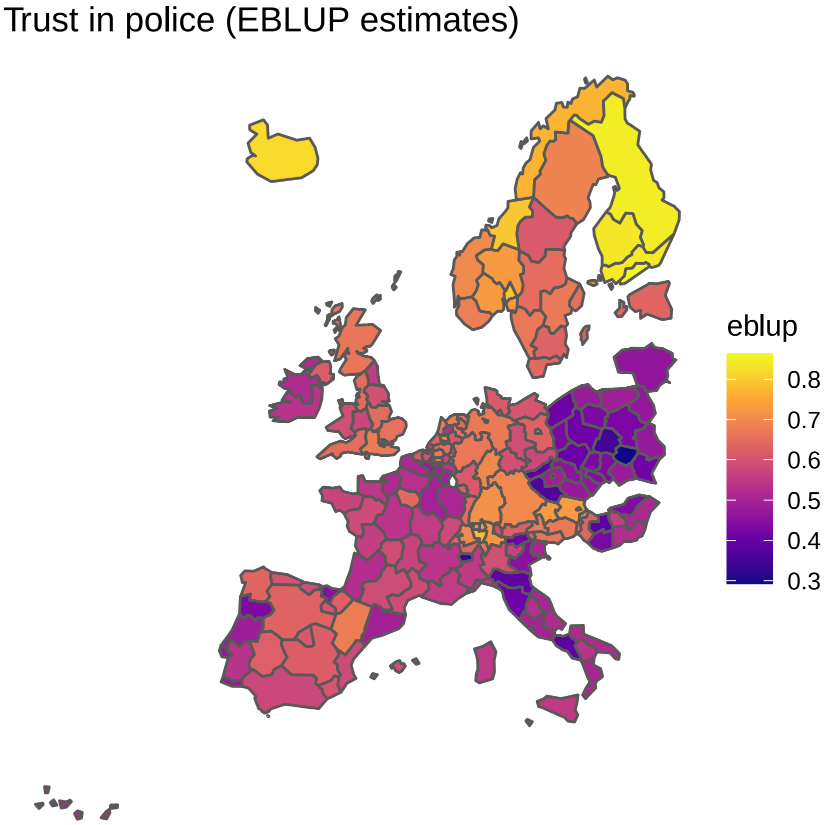
```
st_crs(nuts) <- 15752 #change CRS to ED79 (EPSG:4668 with transformation)
```

Our shapefile of European regions is now ready, but we still need to attach our EBLUP estimates of trust in the police before we can visualize them. We can use function from the `dplyr` package seen above to do this.

```
geodata <- nuts %>%                  #set dataset to use
  rename("domain" = "NUTS_ID") %>%   #rename column of NUTS
  left_join(pop, by = "domain") %>%  #merge area-level estimates
  filter(!is.na(eblup))              #remove areas without estimate
```

The last step is to map our small area estimates of trust in police in Europe. We can use various functions from `ggplot2` and `sf` to produce a clear and informative visual output.

```
ggplot(data = geodata) +                       #set shapefile to use
  ggtitle("Trust in police (EBLUP estimates)")+#change title
  geom_sf(aes(fill = eblup)) +                 #draw map of EBLUPs
  theme_void() +                               #use empty background
  scale_fill_viridis_c(option = "plasma")      #blue-to-yellow scale
```



The map shows how trust is the police is generally larger in the North of Europe, especially in regions of Finland, Norway and Iceland. Some regions in Central Europe (e.g., Switzerland, Germany and Austria) also have large estimates of trust in the police. On the opposite side, many regions in Eastern Europe, especially in Poland, Lithuania and Hungary, have much lower levels of trust in the police. There are also regions in South Europe where the estimates of trust in police are low, such as some regions in Italy and Portugal. We also observe important heterogeneity within some countries. In Spain, for example, our estimates of trust in the police show much lower scores in the Basque Country and Catalonia than in the rest of the country.

4.8 Computing the RRMSE of EBLUP estimates

One of the main benefits of using SAE is that we can estimate the reliability of each small area estimate. This allows us to check if using model-based SAE is useful to improve the reliability of our estimates in all areas, or whether model-based estimates are reliable in some areas but not others.

We will calculate the RRMSE of our EBLUP estimates following an analytical approximation as in Datta & Lahiri (2000), and then compare the measures of reliability of our model-based estimates with the CV of the direct estimates (Rao & Molina, 2015). We use the `mseFH()` function from `sae` to calculate the MSE of our estimates. The arguments of this functions are the same used before in the `eblupFH()` function.

```
eblup_mse <- mseFH(formula = pop$Direct        ~ pop$fem_p_16  +
                             pop$gdp_eurhab_16 + pop$robb_r_10 +
                             pop$burg_r_10     + pop$medage_16 +
                             pop$he_p_16,
                   vardir  = pop$SD^2,
                   method  = "REML")
```

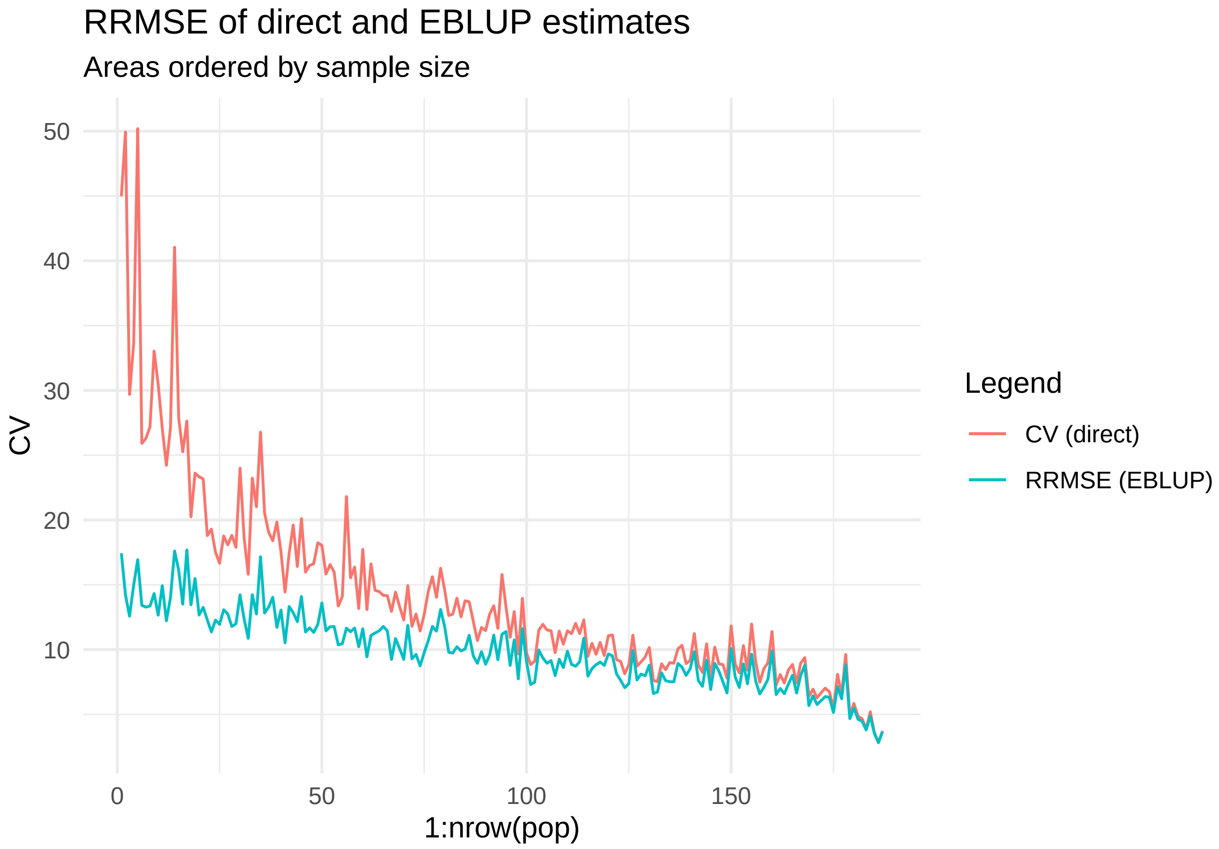
Once we have calculated the MSE of our estimates, we can merge this information with all our area-level data, and use these to calculate the RRMSEs using the `mutate()` function from `dplyr`.

```
pop <- pop %>%                         #set dataset to use
  cbind(eblup_mse$mse) %>%             #merge MSEs of EBLUPs
  rename(mse = "eblup_mse$mse") %>%    #change name of column
  mutate(rrmse = (sqrt(mse)/eblup)*100)#compute RRMSEs
```

4.8.1 Plotting the RRMSE of EBLUP estimates

After that, we can visualize the RRMSE of our EBLUP estimates with the CV of direct estimates to analyze if all our efforts to produce reliable model-based estimates have paid off, and check whether we have successfully improved the reliability of our small area estimates. We will plot the RRMSE of EBLUP estimates and the CV of direct estimates using the `ggplot2` package.

```
pop %>%                                #set dataset to use
  arrange(SampSize) %>%                #order by area sample size
  ggplot() +                           #generate a plot
  geom_line(aes(y     = CV,            #plot line of CV of direct
                x     = 1:nrow(pop),   #visualize one point per area
                color = "darkred")) +  #plot with red line
  geom_line(aes(y     = rrmse,         #plot line of RRMSE of EBLUPs
                x     = 1:nrow(pop),   #visualize one point per area
                color = "steelblue")) +#plot with blue line
  scale_color_discrete(name = "Legend",#add legend and change title
                       labels = c("CV (direct)", "RRMSE (EBLUP)")) +
  ggtitle(label    = "RRMSE of direct and EBLUP estimates",
          subtitle = "Areas ordered by sample size")
```



We had seen before that smaller values of RRMSE indicate a better reliability of estimates, and estimates with a RRMSE smaller than 25% may be considered reliable enough to be used for research, policy making and policing operational decisions. We can see that our EBLUP estimates of trust in the police are more reliable than the direct estimates in every single region, and EBLUP estimates have much higher levels of reliability in those areas where original sample sizes where smaller. All our EBLUP estimates have levels of RRMSE below 25%. This is a sound indicator of reliability in our estimates.

4.9 Model diagnostics

When we use model-based SAE, another key step is to present the diagnostics of our model (Pfeffermann, 2013). Before, we have discussed why regression-based synthetic estimates are generally not accepted to produce estimates (i.e., these have a high risk of bias; Levy, 1979), but we also need to ensure that our EBLUP estimates are not affected by model misspecification. There are different diagnostics that we can use to check the extent to which our estimates may be affected by bias arising from the model. Here we will illustrate this by visualizing a scatter plot of our model-based estimates against the direct estimates. Remember that direct estimates may not be reliable in many areas, but these do not suffer from any type of bias arising from any model (these depend only on the survey sample). In order to check if our model-based estimates suffer from some sort of bias arising from the model, we can compare these with the design-unbiased direct estimates. We will visualize these scatter plots for both our EBLUP and regression-based synthetic estimates against the direct estimates. We expect a much stronger linear association between our EBLUP and direct estimates than between the synthetic and direct estimates, which would indicate that our EBLUP estimates do not suffer from much bias arising from the model.

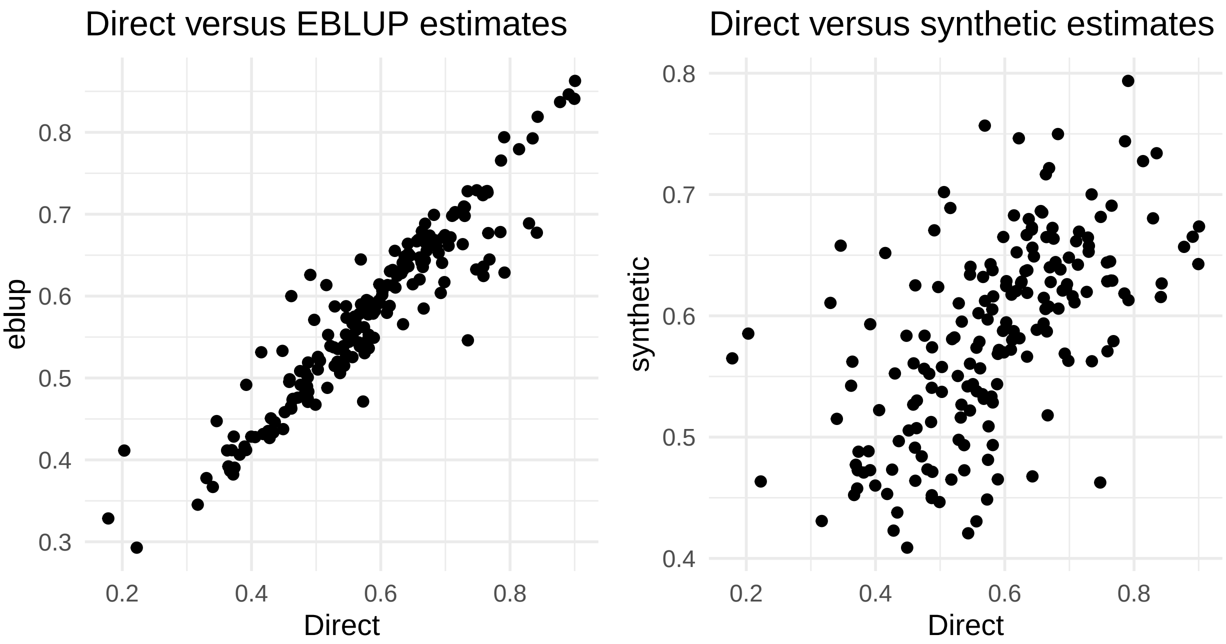
We will use the `ggplot2` library to code our two scatter plots, and visualize them with the assistance of the `gridExtra` package (Auguie, 2017).

```r
library(gridExtra) #load gridExtra package

#plot direct estimates against EBLUPs
dir_vs_EBLUP <- ggplot(pop, aes(x = Direct, y = eblup)) +
                geom_point() +
                ggtitle("Direct versus EBLUP estimates") +
                theme(aspect.ratio = 1)

#plot direct estimates against synthetic estimates
dir_vs_synth <- ggplot(pop, aes(x=Direct, y=synthetic)) +
                geom_point() +
                ggtitle("Direct versus synthetic estimates") +
                theme(aspect.ratio = 1)

#visualize the two plots next to each other
grid.arrange(dir_vs_EBLUP, dir_vs_synth, ncol = 2)
```



There is a much stronger linear association between our EBLUP estimates and the direct estimates than between synthetic estimates and direct estimates. In other words, regression-based synthetic estimates suffer from bias arising from the model, whereas EBLUP estimates are generally unbiased. We can further explore this by analyzing the Spearman's rank correlation (i.e., `cor.test(pop$Direct, pop$eblup, method = "spearman")`), which indicates that the rank correlation between EBLUP and direct estimates equals 0.94 (p-value < 0.001), whereas such correlation is only 0.64 (p-value < 0.001) when we compare direct estimates with regression-based synthetic estimates. We can also calculate the model standardized residuals and present their Normal Q-Q plots in order to check the normality of the residuals (see Buil-Gil, Moretti, et al., 2019 for an example).

5. Final remarks

Research on crime and place is moving toward the study of very detailed levels of analysis, such as street segments and addresses (Groff et al., 2010; Weisburd et al., 2012). In order to analyze the concentration of crime at such detailed level of spatial granularity, it is key to ensure that data are not affected by sampling bias, coverage error or measurement error (Brantingham, 2018). Victimization surveys provide key information to study crime (known and unknown to the police) and trust in the police (Rosenbaum & Lavrakas, 1995; Xie & Baumer, 2019), but these surveys are not designed to produce reliable direct analyses at the increasingly refined spatial scales of the criminology of place. Victimization surveys are generally designed to allow producing precise direct estimates for very large areas, such as countries or states, but sample sizes in smaller areas are usually too small to allow producing direct estimates of adequate precision. Model-based SAE techniques may be used to improve the reliability of small area estimates produced from victimization surveys for place and crime research (e.g., Buelens & Benschop, 2009; Buil-Gil, Medina, et al., 2019). Here we have seen how the area-level EBLUP, which is one of the basic approaches for model-based SAE (Rao & Molina, 2015), may be used to produce reliable small area estimates of trust in the police even when original sample sizes are too small to allow producing reliable direct estimates. Model-based SAE techniques can also be applied to produce estimates of crime in areas (Fay & Diallo, 2015).

In this chapter, we have seen how the ESS records and publishes highly relevant data to analyze crime in Europe, but crime analysts and criminologists may also explore the use of many other surveys that include measures of victimization and trust in the police in other countries. For instance, the NCVS and the CSEW record many questions to study crime and police legitimacy, and SAE can be applied from their data to produce small area estimates of crime and many crime-related measures. The main limitations one may encounter when aiming to apply SAE to data recorded from victimization surveys are *(a)* low-level geographical information may not be published openly, *(b)* difficulty in finding adequate covariates to estimate SAE models, and *(c)* inadequacy of traditional model-based SAE approaches to estimate parameters under non-normal distributions (e.g., Poisson distribution, zero-inflated distributions). With regards to the first issue (i.e., low-level spatial data not being published), researchers sometimes need to request special access to such data via safe rooms. In some cases, the process of requesting access and receiving the data can take months. The second limitation (i.e., access to adequate covariates) is sometimes difficult to solve, since model-based SAE applications frequently rely on public administrations making data open access, which we then use as covariates in our models. The access to high-quality open data is, once again, key for research. Finally, with regards to the limitations of traditional model-based SAE to estimate parameters under non-normal distributions, we can highlight that a new generation of SAE researchers are today investigating the development of more refined techniques to estimate parameters under more complex distributions. Crime and place research will clearly benefit from new techniques that allow estimating parameters of variables with zero-inflated and Poisson distributions that typically define crime victimization data.

Author bio

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 the 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.

Acknowledgments

The author would like to thank Samuel Langton and Angelo Moretti for comments that greatly improved the manuscript.

References

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics*. https://CRAN.R-project.org/package=gridExtra

Braga, A. A., Weisburd, D., & Turchan, B. (2018). Focused deterrence strategies and crime control: An updated systematic review and meta-analysis of the empirical evidence. *Criminology & Public Policy*, *17*(1), 205–250.

Brantingham, P. J. (2018). The logic of data bias and its impact on place-based predictive policing. *Ohio State Journal of Criminal Law*, *15*, 473.

Brunton-Smith, I., & Jackson, J. (2012). Urban fear and its roots in place. In V. Ceccato (Ed.), *The urban fabric of crime and fear* (pp. 55–82). Springer.

Buelens, B., & Benschop, T. (2009). *Small area estimation of violent crime victim rates in the netherlands*. Technical Report DMH-2009-02-03-BBUS, Statistics Netherlands.

Buil-Gil, D., Medina, J., & Shlomo, N. (2019). The geographies of perceived neighbourhood disorder. A small area estimation approach. *Applied Geography*, *109*, 102037.

Buil-Gil, D., Moretti, A., Shlomo, N., & Medina, J. (2019). Worry about crime in europe: A model-based small area estimation from the european social survey. *European Journal of Criminology*, 1477370819845752.

Cao, L., Lai, Y. L., & Zhao, R. (2012). Shades of blue: Confidence in the police in the world. *Journal of Criminal Justice*, *40*(1), 40–49.

Cimentada, J. (2019). *Essurvey: Download data from the european social survey on the fly*. https://cran.r-project.org/web/packages/essurvey/essurvey.pdf

Commonwealth Department of Social Services. (2015). *Survey of disability, ageing and carers, 2012. Modelled estimates for small areas, projected 2015*. Australian Bureau of Statistics, Release 1.

CRAN team. (2020). *RCurl: General network (http/ftp/...) client interface for r*. https://cran.r-project.org/web/packages/RCurl/index.html

D'Alo, M., Di Consiglio, L., & Corazziari, I. (2012). *Small area estimation for victimization data: Case study on the violence against women*. New Techniques; Technologies for Statistics 2012 seminar, EUROSTAT.

Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 613–627.

Fay, R. E., & Diallo, M. S. (2015). *Developmental estimates of subnational crime rates based on the national crime victimization survey*. BJS, Office of Justice Programs.

Fay, R. E., & Diallo, M. S. (2012). Small area estimation alternatives for the national crime victimization survey. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 3742–3756.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873.

Groff, E. R., Weisburd, D., & Yang, S. M. (2010). Is it important to examine crime trends at a local "micro" level?: A longitudinal analysis of street to street variability in crime trajectories. *Journal of Quantitative Criminology*, *26*(1), 7–32.

Groves, R. M., & Cork, D. L. (2008). *Surveying victims: Options for conducting the national crime victimization survey*. The National Academies Press.

Hough, M., Jackson, J., & Bradford, B. (2014). Trust in justice and the legitimacy of legal authorities: Topline findings from a european comparative study. In S. Body-Gendrot, M. Hough, K. Kerezsi, R. Levy, & S. Snacken (Eds.), *The routledge handbook of european criminology* (pp. 243–265). Routledge.

Hummelsheim, D., Hirtenlehner, H., Jackson, J., & Oberwittler, D. (2011). Social insecurities and fear of crime: A cross-national study on the impact of welfare state policies on crime-related anxieties. *European Sociological Review*, *27*(3), 327–345.

Jackson, J., Bradford, B., Stanko, B., & Hohl, K. (2013). *Just authority?: Trust in the police in england and wales*. Routledge.

Kääriäinen, J. T. (2007). Trust in the police in 16 european countries: A multilevel analysis. *European Journal of Criminology*, *4*(4), 409–435.

Lahti, L., Huovari, J., Kainu, M., & Biecek, P. (2020). *Eurostat: Tools for eurostat open data*. https://CRAN.R-project.org/package=eurostat

Levy, P. S. (1979). Small area estimation-synthetic and other procedures, 1968-1978. *Synthetic Estimates for Small Areas: Statistical Workshop Papers*.

Molina, I., & Marhuenda, Y. (2020). *Sae: Small area estimation*. https://CRAN.R-project.org/package=sae

Namazi-Rad, M. R., & Steel, D. (2015). What level of statistical model should we use in small area estimation? *Australian & New Zealand Journal of Statistics*, *57*(2), 275–298.

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

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*(1), 40–68.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation. Second edition*. Wiley.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rosenbaum, D. P., & Lavrakas, P. J. (1995). Self-reports about place: The application of survey and interview methods to the study of small areas. In J. E. Eck & D. Weisburd (Eds.), *Crime and place. Crime prevention studies* (Vol. 4, pp. 285–314). Criminal Justice Press; Police Executive Research Forum.

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), 193–211.

Staubli, S. (2017). *Trusting the police: Comparisons across eastern and western europe*. Transcript Verlag.

Tanton, R., Jones, R., & Lubulwa, G. (2001). *Analyses of the 1998 australian national crime and safety survey*. Character, Impact; Prevention of Crime in Regional Australia Conference.

Weisburd, D., Groff, E. R., & Yang, S. M. (2012). *The criminology of place: Street segments and our understanding of the crime problem*. Oxford University Press.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., 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

Williams, D., Haworth, J., Blangiardo, M., & Cheng, T. (2019). A spatiotemporal bayesian hierarchical approach to investigating patterns of confidence in the police at the neighborhood level. *Geographical Analysis*, *51*(1), 90–110.

Xie, M., & Baumer, E. P. (2019). Neighborhood immigrant concentration and violent crime reporting to the police: A multilevel analysis of data from the national crime victimization survey. *Criminology*, *57*(2), 237–267.