Small area estimation and georeferencing in the context of deficient data - Methodology and applications with data from the Munich municipal statistics

Master thesis of the master program:
Statistics with an economic and social science orientation
European Master in Official Statistics

Ludwig-Maximilians-Universität Munich - Department of Statistics

Anian Rottmüller

supervised by

Prof. Dr. Thomas Augustin

in cooperation with the Statistical Office of the City of Munich

Munich, February 22, 2023

Abstract

In recent years, the need for small-scale statistics has increased rapidly in many areas such as politics or economics, but also in official statistics. However, the data basis for estimation in these small areas is often poor and is further degraded by the influence of deficient data like measurement errors, missing data or confidentiality methods. For such cases, this thesis presents various small area estimators which, in contrast to direct estimators such as the Horvitz-Thompson estimator, are intended to provide precise estimates with as low variance as possible even in small areas with few sample elements. For this purpose, the principle of borrowing strength is applied, in which, in addition to auxiliary variables, information from other areas is also incorporated into the estimation. By using random effects in a mixed model, the Battese-Harter-Fuller estimator and the Fay-Herriot estimator convince with particularly positive properties in the context of small area estimation.

The basis for spatial small area estimation is data that are linked to so-called georeferences, which assign a spatial reference to the individual observations. Therefore, georeferencing is also examined more closely in this thesis as a foundation, whereby the use of grids in particular offers a wide range of analysis possibilities in official statistics. For the practical analysis, a population dataset from the official statistics of the city of Munich is used, which is also georeferenced. To estimate the duration of residence, the Horvitz-Thompson estimator proves to be unusable due to its large variance. Instead, the indirect post-stratified estimator and the model-based Battese-Harter-Fuller and Fay-Herriot estimators are convincing. In practice, these provide better results than the direct estimators even when estimating under deficient data such as missing data or measurement errors.

Contents

Chapter 1

The importance of small area statistics

The average purchase value for a square metre of land in Germany is 142.70 euros in 2019. Although this statement contains an accurate value about the average land price in Germany, it is nevertheless of little help when calculating the cost of buying land, as this is one of the issues with the greatest regional fluctuations throughout Germany. Here it helps to look at land prices on a smaller scale. Even at the federal state level, one can see clear differences between the areas. For example, the average purchase price in the city states of Berlin and Hamburg is over 1,000 euros per square metre. But there are also large discrepancies in the other federal states. For example, the price per square metre in the territorial states varies between just under 37 euros in Thuringia and 211 euros in Bavaria. This breakdown already provides a more accurate picture of land prices. The situation becomes even clearer if one makes even finer breakdowns and looks at the situation at the district level. The data for this are also included in the database of the Regionaldatenbank Germany (Statistische Ämter des Bundes und der Länder 2022). For this comparison, the state of Bavaria is considered, with its average price per square metre of 211 euros. The most expensive prices for a square metre of building land are found in the city of Munich at 2540 euros and the district of Munich at just under 1250 euros, followed by the district of Starnberg at 965 euros. All these areas are in Upper Bavaria, where on the other hand there are also districts, such as Mühldorf am Inn or Altötting, which are less than 100 kilometres away from Munich and where the price per square metre is just over 100 euros. It gets even more extreme when one looks at districts in Franconia, another part of Bavaria, such as Kulmbach or Wunsiedel, where the purchase price for a square metre of building land is around 20 euros. All these differences only become visible when one looks at values on a smaller level. If only Germany or the federal states were considered, these discrepancies would never have come to light. This aspect becomes even more apparent if one were to conduct even smaller-scale studies. However, the lowest level of the data set considered in the Regionaldatenbank Germany is the district level.

This example has already shown how important it can be in many areas to examine facts and statistical analyses even at smaller spatial dimensions. In order to be able to carry out small-scale studies at all, it is important that the data have a spatial reference, so-called georeferences, because this spatial reference often contributes to a better understanding of statistical facts in general as well as in official statistics. Many of the surveyed facts show clear regional differences, as was already evident in the case of purchase prices for building land. This provides an important basis for regional planning and considerations (Brenner 2015: p. 47). Georeferenced data can be used to make small-scale policies that can guide current and future generations in the respective areas. Particularly for labour market, education, medical and socio-economic research, comprehensive and comparable geodata of high quality that can be evaluated on a small scale are required (Schnorr-Bäcker 2012: p. 564).

Data with georeferences and small-area statistical analyses are also the central topic of this master's thesis within the EMOS programme. Thus, besides georeferencing itself, small area estimation is the core of this paper. Special attention is paid to these topics with regard to deficient data. Therefore, after this introduction, the second chapter gives a general insight into deficient data with the different types of it. Then, in the third chapter, georeferencing is presented as a method for giving data a location reference. Here, the development of georeferencing in official statistics and with regard to Big Data in recent years is also discussed. After the data can be assigned to a location with the help of georeferences, the fourth chapter gives an introduction in small area estimation, with which statistical analyses can also be carried out on a small scale. The importance of this was already given at the beginning of this thesis. The fifth chapter goes even deeper and presents the Battese-Harter-Fuller estimator and the Fay-Herriot estimator, the two central estimation methods of small area estimation. In the sixth chapter, georeferencing and small area estimation are examined with regard to deficient data, before in the seventh chapter, population data from the statistical office of the city of Munich is analysed. Here, small area estimation and the methods of official statistics with regard to georeferencing are examined in practice and the effects of deficient data are tested. In the last chapter, a conclusion and an outlook are given.

Chapter 2

Deficient data

Coarse data, rounded data, censored data, missing data and so on. All these terms are in some way related to deficient data and show how multifaceted this aspect is. Therefore, this chapter will go into more detail about deficient data and its different types, as this knowledge will be needed later in the context of georeferencing and small area estimation. While in the case of missing data every value is either perfectly observed or completely unknown, there are many other situations where the data is neither completely missing nor perfectly available. These cases are called coarse data. Here, only a subset of the full sample space in which the true data lies is observed. This includes, among others, rounding or censoring (Heitjan, Rubin 1991: p. 2244 f.). All these mentioned concepts will be gradually introduced in more detail in the course of the chapter, starting with the missing data.

2.1 Missing data

Missing data is certainly one of the main problems when thinking about deficient data, which is why it is started with. Missing data are for example composed of refusals to answer by persons selected for the survey, denials to individual questions or defective measuring devices. Since missing data is a very diverse phenomenon, there is no single definition for it. In general terms, missing values are unobserved data values of variables of scientific interest that are assumed to exist but cannot be formed from other observed values by deterministic methods (Kleinke et al. 2020: p. 1 f.). In surveys, one can also distinguish between item nonresponse and unit nonresponse. In item nonresponse, the values for one or more variables are missing for certain respondents. In the case of unit nonresponse, on the other hand, no single value is observed for an entire respondent, for example due to refusal to participate. Another case that can occur in panel studies is panel attrition. Here, participants are observed in the first waves and drop out at some point, which is why their data are missing in later waves (Kleinke et al. 2020: p. 2). In order to better deal with missing data, one needs to look at the different mechanisms underlying the missing. Since these mechanisms have a very central importance in the treatment of missing values, they are also briefly presented here.

The different missing data mechanisms

In order to be able to describe the missing data mechanisms more precisely, a certain notation is required. Y denotes the complete data set. The index i stands for the i-th observation and the index j stands for the j-th variable, so that y_{ij} indicates the value of the variable Y_j for the i-th observation. Based on this, one can generate the missing data matrix R, which takes the value $r_{ij} = 1$ if y_{ij} is a valid value and $r_{ij} = 0$ if y_{ij} is missing (Little, Rubin 2002: p. 4). According to Rubin's theory, these missing data variables should be considered as random variables to which a distribution can be assigned. In addition, ϕ stands for the unknown parameters, so that the missing data mechanism can be written as a conditional distribution of R given Y and ϕ, i.e. $f(R \mid Y, \phi)$. Y can be further split into the observable part Y_{obs} and the missing components Y_{miss} (Little, Rubin 2002: p. 11 f.). Based on this notation, three different missing data mechanisms can be distinguished.

Missing Completely at Random (MCAR)

In the first mechanism, the data is missing completely at random (MCAR). For this, the probability of R must assume the same value for each ϕ, independent of Y. This means that the distribution of R does not depend on the values of Y, whether Y is observed or unobserved. Formally speaking, the following must apply to MCAR:

$$f(R \mid Y, \phi) \;=\; f(R \mid Y_{obs}, Y_{miss}, \phi) \;=\; f(R \mid \phi) \text{ for all Y, } \phi$$

(Kleinke et al. 2020: p. 26).
For instance, a simple example of MCAR is a survey in which age and income are collected. Age is collected in full, but due to a technical defect, about one fifth of the values for the income variable are missing. Here, the missing is MCAR because the probability of a valid value for income is the same for all respondents, regardless of age and observed income values.

Missing at Random (MAR)

The second missing data mechanism is missing at random (MAR). The data are missing at random if the distribution of R, conditional on the observed values, is independent of the unobserved data. That is, the function of R assumes the same probability for each possible value of the unobserved data, which can be written formally as:

$$f(R \mid Y, \phi) \;=\; f(R \mid Y_{obs}, Y_{miss}, \phi) \;=\; f(R \mid Y_{obs}, \phi) \text{ for all } Y_{miss},\, \phi$$

(Kleinke et al. 2020: p. 28).
To give an example of MAR, one considers again a data set with the variables age and income, assuming that the two characteristics are negatively correlated. If younger respondents tend not to answer the income question, the missing values for income are MAR.

Not Missing at Random (NMAR)

The last missing data mechanism is not missing at random (NMAR). This case occurs when the missing data is neither MCAR nor MAR. In NMAR, the distribution of R also depends on unobserved values Y_{miss}. The following expression cannot be further simplified under not missing at random:

$$f(R \mid Y, \phi) \;=\; f(R \mid Y_{obs}, Y_{miss}, \phi)$$

(Kleinke et al. 2020: p. 30).
In the example with the two characteristics age and income, not missing at random is present if, in the case of persons of the same age, those with the higher income tend not to answer.

Ignorability of missing data

In order to be able to deal with the missing data, it is important to know in which cases the mechanism can be ignored without any problems. For this, a huge dataset with data pairs $(x_i,\ y_i)$ is considered. In the first case, some of the values of the first variable x are removed completely at random. Since MCAR is present here, the bivariate distribution of $(x_i,\ y_i)$ does not differ systematically between the observed and real data set. This means that the standard statistical analysis techniques can be used without any problems if the model assumptions are approximately correct. For example, a linear regression estimator would not differ between the two data sets with and without missing x-values (Kleinke et al. 2020: p. 27).

In the next case, the same scenario is considered under MAR. This time, the x-values are not deleted completely randomly, but depending on the associated y-value. Here, both the estimated marginal distribution of the y values whose associated x values are observed and the estimated marginal distribution of the observed x values differ from the associated distributions of the fully observed x and y variables. Thus, even estimators like the mean or variance, are biased when based on the fully observed pairs $(x_i,\ y_i)$. The regression estimators for the link between x and y, on the other hand, are not biased and do not differ systematically between the observed and true data if MAR holds (Kleinke et al. 2020: p. 28 f.).

To illustrate the case of NMAR, again a data set with the value pairs $(x_i,\ y_i)$ is considered. In this case, the x-values of some observations are again deleted, but this time the probability of being deleted depends on x itself. Under these conditions, not only the distributions of the variables differ between the observed and complete data set. The regression relationship is also estimated differently for the data with MNAR than for the true data. To deal with this, good knowledge of the missing data mechanism is needed in the NMAR case (Kleinke et al. 2020: p. 30 f.).

This knowledge is helpful in order to then be able to deal specifically with the missing data, whereby the type of analysis must also always be taken into account. For example, it became clear in this section that linear regression estimators based on the observed data are unbiased to the estimates on the complete data set under MCAR and MAR. Under NMAR, on the other hand, there are already differences between these two estimates. This concludes the section on missing data and other types of deficient data are presented in the remainder of this chapter.

2.2 Censoring

Another type of deficient data, similar in some ways to missing data, is censored data. Censored data result from incomplete data in longitudinal studies and are frequently encountered in survival and reliability analysis (Islam, Chowdhury 2017: p. 6 f.). In the previous section on missing data, panel attrition was already mentioned. This could also be considered as censoring. The problem with censoring is that not all the information is available about a subject, but that only some parts of it are known. As a result, it happens that one does not know the exact time at which an event of interest occurs. The various types of censoring differ according to what information is available about the subjects.

Right censoring occurs in a large number of studies, which is why it is subdivided into different types. In Type I censoring, a predetermined time t_0 is set for the study. The exact lifetime of an individual is only known if it is smaller than t_0. Otherwise it is censored (Islam, Chowdhury 2017: p. 195). For example, in a clinical trial with sick patients, the trial may only go on for a certain period of time and not all patients will be cured within that period. Their duration to cure would then be censored.

In Type II censoring, only the values of the smallest r units in a sample are observed. After the event of interest has occurred in these r observations, the study is terminated. The values of the other $n - r$ units are censored. For these units, one only knows that the event did not occur until the time of censoring (Islam, Chowdhury 2017: p. 195). An example of this would be a clinical trial with 500 sick patients. This is stopped when 300 patients have been cured. For the remaining 200 patients, the time to cure would be censored.

Another type of right censoring is random censoring. This occurs when units are censored for reasons outside the study plan and beyond the control of the director of studies (Islam, Chowdhury 2017: p. 195). For example, in random censoring, units may be censored that do not show up for the examination and other subjects no longer have any value because the study has been discontinued.

A further form of censoring is left censoring. Left censoring occurs in trials where for some units the event of interest is already present at the beginning. Here, all subjects are followed until the event has occurred and the time from that is of interest. However, if the event has already occurred at the beginning of the study, the time to occurrence is unknown and the data is left-censored (Islam, Chowdhury 2017: p. 195 f.). An example of this would be a study to measure after how many days after a positive Covid test the patients are asymptomatic again and how long they were symptomatic overall. However, people who have an asymptomatic infection never have symptoms and the time cannot be measured. Other people had symptoms before the test but are symptom-free by the time of the test. These people are left-censored. Another form of censoring would be interval censoring. In practice, however, this occurs much less frequently than right and left censoring and is only mentioned here. Instead, it is looked at the role of censoring in official statistics.

Censoring also plays an important role in official statistics. Often enormously large data sets are considered over a long period of time, and often people are not considered over the whole period because, for example, they die in the middle of it. The Munich population data analysed later is also a longitudinal data set. In this thesis, only one point in time is considered, but the data set of the statistical office has included all residents of Munich for more than 40 years and, in addition to births and deaths, censoring also occurs due to people moving in and out.

2.3 Measurement errors

The next type of deficient data is measurement error. Measurement error is one of the errors that can occur on the path of data collection from sample selection to the cleaned data set and is therefore also part of the Total Survey Error concept (TSE), in which the quality of the data is assessed. Since small area estimation is later also an estimation procedure, it is equally affected by the various errors of the TSE concept, as these lead to deviations between the parameter estimate and the true population value (Faulbaum 2014: p. 439 f.). The TSE concept is divided into two branches. The first contains the sampling errors, while the second branch includes the non-sampling errors. A further distinction can also be made between the representativeness problem and the measurement problem. For this section, however, only the branch with the non-sampling errors, i.e. the measurement problem, is relevant (Groves, Lyberg 2010: p. 856).

Faulbaum (2014: p. 444 f.) also includes in the branch of non-sampling errors the nonobservation errors, such as unit or item nonresponse, which were already considered in the previous section. In the present case, measurement errors in the broader sense are all those cases that Faulbaum (2014: p. 448 ff.) refers to as observation errors. These observation errors consist of measurement errors, processing errors and technical errors. Measurement error in the narrower sense refers to the difference between the true measurement and the observed or recorded value (Groves, Lyberg 2010: p. 855). Within the framework of classical test theory, quantitative variables can be decomposed. Here, the observed value of a variable χ is divided into a true value or score τ and a measurement error ξ:

$$\chi = \tau + \xi.$$

Thus, an observed variable is considered to be the sum of a true variable and an error variable. Rearranging the formula, the measurement error is the difference between the measured value and the true value:

$$\xi = \chi - \tau.$$

This decomposition always works with latent variables, since both the true value and the error variable are unobservable. In surveys, possible sources of error include for instance the behaviour of the interviewer, the interview situation, the design of the questions and the questionnaire or the type of interview (Faulbaum 2014: p. 448 f.).

Processing errors include, for example, input errors or editing errors. These are also included in the difference between the observed and the true value. Technical errors are also gaining in importance due to new digital methods of data collection. A failure of the data collection equipment or erroneous software can lead to problems here (Faulbaum 2014: p. 449).

2.4 Rounding

In addition to measurement error, rounding error must also be considered in the context of deficient data. When collecting data in practice, there are often cases where rounding of continuous variables cannot be prevented. Reasons for rounding include for instance the precision of the measuring device or the limitations of the storage mechanism (Zhao, Bai 2012: p. 895 f.). One example where one does not necessarily think of rounding is the age. In the vast majority of cases, age is given in years, but one could also give months and days, whereby there are no lower limits and hours and minutes, for example, can also be included. For instance, 22 years of age gives away a lot of information. The person under consideration could have just turned 22, but could also be about to turn 23 in a few days. Due to rounding, almost a year of information can be lost here. This discretises continuous values, so to say. Another type of rounding is given when measurements are taken several times and then the mean value is used. Until now, rounding errors were often ignored, especially in smaller data sets. However, this must not be done for data sets with many observations, which are becoming more and more frequent due to the Big Data phenomenon (Ushakov, Ushakov 2018: p. 770 f.). For example, it has been shown that with a sufficiently large sample size, any null hypothesis is rejected with a probability of 1 when rounded data are used. Therefore more precise results are desired.

2.5 Outliers

Later, in the small area methods, model-based estimators are presented which follow a model or a distribution. In this context, outliers can be considered as deficient. Outliers are regarded as strong deviations from the assumed (normal) distribution. Outliers are considered to be correctly recorded and it cannot be expected that there will be no further such deviations in the non-sampled part of the population. Due to the strong deviations from the assumed distribution, outliers are seen as particularly influential observations. It is presumed that a small part thus originates from a distribution other than the assumed one (Sinha, Rao 2009: p. 382).

2.6 Confidentiality methods

Another aspect that must not be forgotten in this work on the topic of deficient data and which is of great importance here, especially in official statistics, is statistical confidentiality methods by which the data are changed. The fifth principle of the European Statistics Code of Practice deals explicitly with statistical confidentiality and data protection. This is to ensure that data are handled securely and confidentially in all circumstances and that information from respondents is only used for statistical purposes. The principle also implies that strict confidentiality procedures are put in place before data are shared with external users for research purposes. The following section examines the extent to which certain of these procedures alter the data and thereby turn it into deficient data (Eurostat 2018: p. 12). There are different levels of anonymity and the higher the level of anonymity, the more the data is changed. This can lead to a reduction of analytical completeness or a loss of analytical validity (Duncan et al. 2011: p. 46 f.). A distinction can be made between methods that protect tabular data and methods that take care of microdata. Since the population data of the Munich municipal statistics are microdata later in this thesis, methods for tabular data are briefly considered at the beginning before the methods for the protection of microdata are dealt with in more detail.

2.6.1 Methods for protecting tabular data

Statistical tables are used to publish aggregated information classified by categories. Tables become problematic in terms of data protection if low frequencies or small numbers occur and conclusions can then easily be drawn about individuals. To prevent this, there are various protective measures, the aim of which is to protect the individuals while keeping the loss of information to a minimum. Cells where the original value of respondents can be easily inferred are also called risky cells. There are various approaches, such as the dominance rule, the posterior ambiguity rule or the n-rule, which identify the individual cells as risky or not (Duncan et al. 2011: p. 65 ff.). However, these individual approaches will not be discussed in more detail here. Instead, methods that attempt to eliminate the risky cells will be looked at. The first and most popular technique for protecting the confidentiality of tabular data is cell suppression. Here, risky information is protected by hiding the values of some cells. The suppressed values are then usually replaced by a fixed symbol. The risky cells are replaced right at the beginning and are called primary suppressions. However, since the marginal frequencies are often also given in tables, it is not sufficient to carry out only primary suppressions. Other cells must also be found and suppressed to prevent identification, which is called secondary or complementary suppressions. The problem of suppressing values with minimum loss of information is known as cell suppression problem and requires difficult optimization (Duncan et al. 2011: p. 78 ff.). After the cell suppressions have been performed, the data are deficient because whole values are missing for the cells in consideration.

The second method of protecting statistical tables is interval publication, where for risky cells not the exact value is published, but an interval that can vary in length. Initially, interval publication was also known as the partial cell suppression method. Although this method does not produce any missing values, the interval representation nevertheless causes some information to be lost from the raw data (Duncan et al. 2011: p. 81 f.). The next technique is controlled rounding. The difference between this and the previous two methods is that controlled rounding publishes one single value for each cell instead of a missing value or an interval. The user is informed that this is an approximate value and can then decide for themselves whether to use it or create an interval with it. In controlled rounding, a base number r_i is specified for each cell i, for example $r_i = 5$, where each value of the table is rounded to the corresponding base number. This is deficient data as the original values are not published and the loss of information can be seen as the difference between the raw and the published value. The cell perturbation or partial controlled rounding method is based on controlled rounding. Here, a value is selected for the risky cells that lies between the lower and upper rounding limits of the respective cell (Duncan et al. 2011: p. 82 ff.). A very old method to protect data in tables is table redesign. Here, rows or columns that contain too many risky cells are merged. This creates cells that contain aggregated data from multiple respondents and small cell counts can be prevented (Duncan et al. 2011: p. 87). This selection of methods presented aim to publish statistical tables where the data is well protected and safe from data snoopers. As a result, the original data is always altered and there is a loss of information. In the following, methods that are intended to protect microdata are discussed.

2.6.2 Methods for protecting microdata

In contrast to tables, microdata contain values of individuals or individual objects. As a result, they comprise enormous analytical potential, but are also often very difficult to protect. This high analytical potential of microdata in official statistics results, among other things, from large sample sizes and an obligation to provide information, which makes them an important building block of society and political decisions as well as a basis of much research work. It should also be explicitly mentioned here that microdata play an important role in small area estimation as they often contain geographical information, even about small areas. In many of these cases, microdata cannot simply be replaced by aggregated data, despite the high need for protection, so protective measures that modify the data are needed before publication (Duncan et al. 2011: p. 95 ff.). Microdata are not protected at all if they contain direct identifiers, such as id numbers. In addition, the data is very poorly protected if it contains close identifiers, such as names or addresses. These identifiers should be removed from the dataset before publication. It is also important to ensure that the dataset no longer contains identifiers that allow conclusions to be drawn about the individuals with the help of other datasets. In the context of small area estimation, it is also relevant to mention that geographical information can easily be used as identifiers. This aspect will be considered in more detail in the chapters on georeferencing and small area estimation (Duncan et al. 2011: p. 99 ff.). In order to fulfil the points mentioned so far in this section and provide useful and secure data, there are two possibilities. Restricted access to the data requires that only certain licensed persons have access to the data. Restricted data, on the other hand, usually does not restrict access, but transforms the risky source data into safe data or restricted microdata with an extremely low disclosure risk. Confidentiality and security protection can also be established through a combination of restricted access and safe data. In the context of the deficient data, the restricted data will be examined in more detail in the following, since the original data are changed by the various possible methods (Duncan et al. 2011: p. 105 ff.).

Matrix masking

The process of creating restricted microdata can be represented by matrices and is called matrix masking. Here, the original source data of n units to p features are considered as a matrix X of dimension n x p. The mask transformation process transforms the source data X into the masked data Y. This process can be described by the expression $Y = AXB + C$, where A describes the premultiplication of X with the matrix A, which changes the rows. The columns of X are changed by the postmultiplication of X with B. In addition, the matrix C can be added, which directly changes the values of X. The specification of the matrices A, B and C affects both the data utility and the disclosure risk (Duncan et al. 2011: p. 109 f.). This matrix notation helps for many methods to describe the changes of the original microdata.

Suppression methods

The first masking method is masking trough suppression. This technique suppresses certain features of the data set in order to reduce the disclosure risk. A distinction is made between two types of suppression. With record suppression, individual data records are deleted. Here the focus is on premultiplying the source data with a matrix A, which is the identity matrix I with zeros in the places to be suppressed. For example, respondents with income above a certain risky value can be removed from the dataset. With attribute suppression, on the other hand, individual variables, such as the date of birth, are deleted, as with these variables there is an increased risk of identifying individuals with the help of further data sets. For this, X is post-multiplied by the matrix B, where B corresponds to the identity matrix in which the variables to be suppressed are zeros. Geographical information is a feature that is often suppressed because the risk of identification is too high below a certain population size. A more complicated method based on this is local suppression. Here, certain values of individual attributes are suppressed for combinations that have only few observations and are therefore exposed to a high disclosure risk (Duncan et al. 2011: p. 110 f.). The missing values result in deficient data due to record suppression as well as attribute suppression and also local suppression.

Noise addition

Another way to restrict the microdata is via noise addition. In this process, the original value x is perturbed by adding noise in form of a random variable ϵ. This gives the masked value of y from the expression $y = x + \epsilon$. The mean value of ϵ is assumed to be zero so that no bias is added. The variance of ϵ controls the extent of perturbation. By adding the noise to the key variables, it becomes eminently more difficult for data snoopers to link the dataset to other datasets for identification (Duncan et al. 2011: p. 112 ff.). Since the original source data is no longer available when noise is added, the masked data is also deficient data.

Data swapping

Another very multifaceted possibility for the restriction of microdata is data swapping. In general, this method swaps certain values or fields from one unit with those in another unit. The data matrix X of the source data with the dimension n x p becomes the n x p matrix M by swapping. Data swapping reduces the disclosure risk as the data snooper can no longer be sure whether the linkage is correct, while at the same time the data utility for many applications does not decrease to the same extent as for example with the previous methods, since many key figures, such as the arithmetic mean of features in the published matrix M, remain unchanged compared to the original data. Data swapping can also be used in conjunction with other disclosure limitation techniques. At the beginning of the data swapping process, the swapping candidates are selected. In some data sets this can be all units, but most of the time the number of swapping candidates is limited. Candidates are those units that can be easily identified in combination with other data sets. Then the swapping keys are determined. These are the features that are used for matching the records to be swapped. At this point, domain specific knowledge is needed to define the swapping keys. Often attributes are keys when it is assumed that data snoopers have a high knowledge of that characteristic. The swapping attribute is the feature that is exchanged between the swapping candidate and the swapping partner during swapping. The values of the swapping partners for the key variables should be as similar as possible to those of the swapping candidates (Duncan et al. 2011: p. 114 ff.). Since the data is also changed during swapping, deficient data is created here as well.

Sampling and Aggregation

Microdata can also be protected by publishing only a part of it. This is done by masking through subsampling, where the masked data matrix Y is a sample of the source data X. This method is often used in the preparation of public-use microdata. Here, data snoopers can never be sure if linkages to other datasets are correct because they do not know if the object under consideration is in the masked dataset. Subsampling also works well in combination with other disclosure limitation methods (Duncan et al. 2011: p. 118 f.). One of these methods could also be masking through aggregation. This technique is somewhat similar to the table redesign method for protecting tabular data. Aggregation coarsens the data by merging the attributes of several units or by statistically combining the values of different attributes. A distinction is made between global recoding and topcoding. In the former method, several categories of a characteristic are combined into less specific categories. In this characteristic, all units of the dataset are merged from narrower categories into broader categories. For example, geographical information about a county may be merged into categories about a state. This considerably reduces the disclosure risk. However, the utility of the data decreases due to the loss of more precise information. Topcoding is an extension of global recoding. Here, all characteristic values above a certain threshold value are combined, for example, to the conditional median or mean. In the area of aggregation, microaggregation also plays a role. The values of particularly risky attributes are divided into clusters and an average or median value is used for each unit within the cluster (Duncan et al. 2011: p. 119 f.).

Synthetic microdata

All these microdata protection methods presented in this section 2.6.2 are data-conditioned methods and turned source data into masked data. In the approach of generating a synthetic data set, on the other hand, it is assumed that the source data are the realisations of a statistical model. This approach of generating a synthetic dataset is simple in theory, but requires some effort in implementation. First, the distribution of the variables in the data set is estimated from the original data X. This distribution is the basis of a model that is used to generate samples for the synthetic data that replace the original data. This reduces the disclosure risk of synthetic data, as there is no direct link between the source data and the published synthetic data (Duncan et al. 2011: p. 120 ff.).

The methods presented are a selection of techniques that attempt to protect tabular data and microdata so that the fifth principle of the Code of Practice for European Statistics on confidentiality and data protection can be met, but the data can still be used for research purposes, among others. The methods for the protection of microdata were dealt with in more detail, as they play an important role in official statistics. In addition, the practical part of this thesis analyses a microdata set of Munich population data from the statistical office in Munich, for which confidentiality methods were also applied prior to release, which will be discussed in more detail later when the data set is presented.

The chapter on deficient data is concluded with the secrecy methods. It has become clear for what many reasons data does not always correspond to what is desired. This chapter on deficient data also forms a basis for the further course, as deficient data also have an impact on georeferencing and small area estimation. The next chapter introduces georeferencing.

Chapter 3

Georeferencing with regard to official statistics

One way to take spatial and geographical aspects into account in statistical analysis is georeferencing. This chapter sets out what georeferencing is in general before looking at it more closely in a statistical context and presenting its relevance and development in official statistics in recent years, also with regard to Big Data.

3.1 General aspects of georeferencing

First of all, it should be clarified what georeferencing actually is. The term georeferencing basically covers all techniques and methods that serve to the unique identification of geographical objects (Hackeloeer et al. 2014: p. 61). It can also be considered as the first step or stage of spatial data analysis (Ribeiro et al. 2014: p. 2). The aim here is to place maps in a geospatial coordinate system by assigning coordinate values. This is necessary, among other things, if digital maps lack a reference to a geospatial coordinate system (Spektrum der Wissenschaft 2023). In relation to data, georeferencing refers to the assignment of spatial reference information to a spatial data set. Georeferenced data are therefore map data that are related to geographical reference points. The data, e.g. measurement data, are thus linked to coordinates. This plays an important role, for instance, in computer cartography, remote sensing and geoinformation systems (Bundesverband Geothermie 2020). A term that is also frequently used in the course of georeferencing is geocoding. Geocoding refers to the actual transformation step necessary to convert data of various georeferencing into a desired reference system (Professur für Geodäsie und Geoinformatik 2001). It can also be argued that data and information objects, such as datasets, text documents, maps, photographs or imagery are referenced to their proper locations on earth. Most of such objects derive from observations or measurements and are naturally georeferencable, as human activities are tied to the near-surface of the earth. Geospatial referencing originally comes from marine navigation. While approximate position information is usually sufficient on land, precise information on longitude and latitude is essential at sea in order to navigate the desired route (Hastings, Hill 2018: p. 1616 f.).

Basic terms of georeferencing

Now that a general introduction to georeferencing has been given, some basic terminology of it is introduced below. In the previous paragraph, the term geographical object has already been used in some points. A *geographical object* can be any type of object or structure that can be linked to a geographical location, such as points of interest, places, roads, buildings or agricultural areas. A *geographical location* is in turn defined as a unit that represents a spatial reference. One can distinguish between different spatial dimensions of geographical locations. For example, points are zero-dimensional, lines are one-dimensional, surfaces are two-dimensional and bodies are three-dimensional (Hackeloeer et al. 2014: p. 61). Another distinction can be made between *formal* and *informal* georeferencing. While formal georeferencing assigns exact spatial coordinates to data and information, informal georeferencing provides colloquial references to existing geographic objects such as street names. In formal georeferencing, *longitude* and *latitude* are often used as coordinates, which depend on the definition of the centre of the earth. In informal georeferencing without exact coordinates, it can happen that several names exist for the same geographical location. In order to translate between formal and informal georeferences, so-called *gazetteers* are used (Hastings, Hill 2018: p. 1616). To reference geographical objects, *geographical*, *topological* and *semantic information* can be used. The term *matching* refers to the process of identifying geographical objects and assigning them to geographical locations. For this purpose, georeferencing usually uses a *geodetic reference system* as for instance WGS-84, which consists of a standard coordinate system for the earth, a reference ellipsoid and a geoid defining a nominal sea level. In matching, it can happen that data from different maps with different reference systems are linked with each other. This combination of map data from different sources to create a new map is also called *conflation* (Hackeloeer et al. 2014: p. 61 f.). Now that certain important terms of georeferencing have been clarified, the next section introduces various methods and techniques.

Georeferencing methods and techniques and their classification

In georeferencing there are a variety of techniques and methods, which are in a sense determined by the format of the underlying data and information objects. Generally speaking, georeferencing can be divided into two types, namely vector and raster referencing, in addition to the distinction between formal and informal georeferencing. Vectors are discrete shapes that can be applied to object-like phenomena, such as points, lines and polygons. Rasters, in contrast, are regular grids that are suitable for field-like phenomena (Hastings, Hill 2018: p. 1618).

In vector referencing, the goal is to identify locations that can be described by means of vectors in order to assign the vector to a corresponding element in a digital map. The easiest way to do this is via positioning or map matching where a coordinate pair has to be assigned to a particular map. But in most cases the coordinates have a certain inaccuracy, which is why certain filtering, rectification and validation methods are still used before matching with the map. Vector referencing can be further subdivided into point referencing, location referencing and road network matching. The first of these three terms refers to the unique identification of a geographical point given by its coordinates in a reference frame. When routes consisting of whole sequences of vectors, such as road segments, or polygons defining areas, are identified, one can speak of location referencing. The last of the three terms, road network matching, is about finding correspondences between graphs created by the road networks of different maps (Hackeloeer et al. 2014: p. 62 f.).

Raster referencing addresses the challenge of correlating pixels in raster images with geographic references. Specifically, it is a projection of features between two spatial reference frames, where at least one feature space is comprised of pixel sets from raster image data. In order to assign locations to the objects contained in raster images, methods such as remote sensing, photogrammetry, computer vision, image processing and pattern recognition are used. Since raster images containing geographic areas often originate from satellite or aerial imagery, the raw images are pre-processed and orthorectified to remove distortions caused, for example, by the camera angle. Afterwards, the processed images are divided into segments, which are then further processed with feature extraction, whereupon a dimensionality reduction is achieved, which allows an easier classification of the objects of the image. In addition, the raster images are linked to digital maps for georeferencing via so-called ground control points. With the help of these ground control points, the pixels of the images can then be referenced with points on the map (Hackeloeer et al. 2014: p. 62 f.).

In the previous paragraph it was already mentioned that three different types of identification properties can be used for referencing geographical objects: geographical, topological and semantic information. According to their properties, the referencing methods and techniques can be classified. Topological properties are preserved in the case of continuous changes of objects. The most relevant objects in digital maps in terms of topological studies are structures that induce a graph, as for example road networks. The ability of topological properties to be independent of geometric variations between two maps simplifies the use of topology to identify geographic entities within various maps, regardless of minor variations in shape, position or other geometric properties (Hackeloeer et al. 2014: p. 64).

Geometrical properties are used to characterise the geometry of an object. Different geometrical parameters can be employed to identify spatial entities. For instance, points are determined by their coordinates, while lines are identified by their geometrical shape. However, one should be careful not to use the geometrical properties of objects as the only identification feature when referencing, as the different databases have different precision in the data (Hackeloeer et al. 2014: p. 64).

In digital maps, all information that is not of a geometrical or topological nature, is assigned to the semantic properties, that define the meaning of an object. Semantic properties can be names of places or points of interest, street numbers, speed limits posted on road segments, among other things. Semantic information is often helpful for the identification of geographic entities, as for example street names found on digital maps rarely do not represent the same street within a small geographic area. However, semantic properties due to, for example, different spellings of streets or places should not be used as the only identifier in georeferencing (Hackeloeer et al. 2014: p. 65).

Now that the different methods of georeferencing and their classification have been introduced, the next section looks at statistical aspects of georeferencing.

3.2 Statistical aspects of georeferencing

Georeferencing plays a role in statistics above all when a location reference is established for units from a data set. The data is often displayed in tables or databases which may contain georeferences in the form of coordinates, place names or even both. In order to deal with statistical data, it is also possible to use place codes instead of place names for administrative or other formally defined areas, like census tracts or postal codes. Data from tables can be georefernced both formally and informally (Hastings, Hill 2018: p. 1619). In each of the following tables 3.1 and 3.2, a simple example is given for a data set without and with georeferencing.

Name	**Sex**	**Profession**	**Monthly income**
Maxi Müller	male	Craftsman	2,500
Franziska Huber	female	Teacher	3,800
Emma Maier	female	Saleswoman	2,100
Thomas Berger	male	Managing director	4,300

Table 3.1: Example of a data set without georeferencing

The same simple sample dataset is used in both tables. While in table 3.1 it does not contain georeferences, in table 3.2 the references are given in the last two columns. The column with the residential address can be seen more as informal georeferencing, whereas the coordinates of this address are formal references in the form of latitude and longitude.

Name	**Sex**	**Profession**	**Monthly income**	**Residential address**	**Geocoordinates**
Maxi Müller	male	Craftsman	2,500	Alexanderpl. 3 10178 Berlin	Lat 52.5215112 Long 13.4150124
Franziska Huber	female	Teacher	3,800	An d. Alster 14 20099 Hamburg	Lat 53.57532 Long 10.01534
Emma Maier	female	Saleswoman	2,100	Hohe Str. 121 50667 Cologne	Lat 50.9353861 Long 6.9561997
Thomas Berger	male	Managing director	4,300	Marienplatz 8 80331 Munich	Lat 48.13743 Long 11.57549

Table 3.2: Example of a data set with georeferences
Coordinates generated with `https://www.gpskoordinaten.de/` (last visit: 22/02/2023)

Using the examples of residential addresses, it is easy to see why there can sometimes be problems with informal georeferencing. Abbreviations are used for the addresses of the first three persons. So, for instance, “Alexanderpl. 3“ and “Alexanderplatz 3“ would refer to the same geographical location. The same applies to “An d. Alster 14“ and “An der Alster 14“ or “Hohe Str. 121“ and “Hohe Straße 121“. The longitudes and latitudes as geocoordinates, on the other hand, are unambiguous as a means of formal georeferencing if they are sufficiently precise.

Another aspect of georeferenced data in statistics is that the spatial reference makes it possible to link data and information from different sources. Thus, data from the most diverse topics, such as society, the environment or the economy, can be joined together (Gebers, Graze 2019: p. 11). In order to be able to implement this, the same basic spatial unit is needed in the different data sets. Geographical rasters are a popular option in statistics for this purpose. These are spatial geometrical reference units that are uniform in size and shape. An example of this would be 100 metre by 100 metre grid cells. Another advantage of geographical rasters is that existing statistics can be expanded to include additional characteristics (Gebers, Graze 2019: p. 12 f.).

In the context of georeferencing, the terms geoinformation and geodata should also be considered in the statistical sense. Geoinformation is defined as information on geographical phenomena directly or indirectly associated with a position related to the earth (Schnorr-Bäcker 2012: p. 563). Thus, geoinformation is location- and space-related data that documents the conditions of a country - be it in the form of coordinates, place names or other criteria (Bundesamt für Landestopografie swisstopo 2023). Geodata, on the other hand, is defined as a digital description of specialised knowledge about geospatial facts and objects, that is, facts and objects associated with a place or a space and their mutual relationships (Schnorr-Bäcker 2012: p. 563 f.). In simple terms, geodata is information to which a spatial location can be assigned. This means that spatial analyses are possible with geodata, even on a small scale. This demonstrates the importance of georeferencing, as the data obtained with it have a high analysis potential. One potential application is small area estimation, which is discussed in great detail in this thesis.

If one looks back at the example tables 3.1 and 3.2, one can see a greater analysis potential in the second table thanks to the spatial references. If the data set contained more observations, the distribution of monthly income in the individual regions in Germany could be examined thanks to the addresses, for example. Now that basic aspects of georeferencing have been considered in a statistical light, the next section will examine georeferencing in the context of official statistics.

3.3 The development of georeferencing in official statistics

One subfield of statistics in which georeferencing now plays an important role is official statistics. Here, georeferenced evaluations expand the range of regional data. In (German) official statistics, regional data are defined as data from smaller regional units below the federal state level, where for a long time the municipal level has been the smallest available unit. The structure of these geodata was mostly given by administrative units. With georeferencing, on the other hand, it is now also possible to obtain data material below the municipal level. But there are a few things to consider, especially from a legal point of view (Brenner 2015: p. 47 f.).

Legal aspects of georeferencing in official statistics

The legal basis for the use of geoinformation in European official statistics is an European Union guideline from 2007. This is the foundation for a common and cross-border spatial data infrastructure, which is known as INSPIRE: Infrastructure for spatial information in Europe. This regulates the unified description and distribution of geodata on the internet for visualisation and downloads (Brenzel, Gebers 2020: p. 49). The development of the unified European spatial data infrastructure follows a step-by-step time schedule. Initially, metadata should be provided for geodata on specific topics. Then all geodata, both new and existing, should gradually be brought into uniform formats, so that from 2021 geodata in all of Europe can be used in the same format (Bundesamt für Kartographie und Geodäsie 2023).

INSPIRE has also contributed to the initiation of measures in many areas of German official statistics in order to be able to provide information on a smaller geographical scale. This led to a renewal of the Federal Statistics Act in August 2013. Here, the basis for geocoding in official statistics was created. Before that, the use of the municipality and the block side as the smallest unit for the regional allocation of the survey characteristics was permissible for a long time. Since the renewal of the Federal Statistics Act, the permanent storage of the spatial reference of statistical data is still not permitted with reference to an address, but with reference to a geographical grid with a minimum grid width of 100 metres, according to paragraph 10. This means that the details of the grid cells in which the coordinates fall should be added before the legally required deletion of the address and coordinate details. Here, the grid cells are coded by an identifier, the grid cell ID. For environmental and economic statistics it is even possible to store the exact coordinates (Brenner 2015: p. 48). A second necessary prerequisite for geocoding in German official statistics is the regulation of the use of official geodata of the surveying administrations of the federal states by federal and state institutions. Since 2019, an agreement on the reciprocal use of official digital geodata between the federal states has regulated the use between the federal state offices in projects without or with the participation of the federal government (Brenzel, Gebers 2020: p. 50).

The most important point in this subsection is that since 2013 it is allowed to permanently store the spatial reference of statistical data in grids of a size of at least 100 by 100 metres, which ensures sufficient protection of the data. The next step is to examine which analyses are carried out as a result of this change in official statistics and how this has resulted in an excellent data base.

Analyses in German official statistics with georeferenced grids

By integrating geographical and statistical data, based on the legal requirements presented in the previous subsection, new information can be obtained in official statistics. In this process, existing data sets are combined and evaluated, so that no additional burden is placed on the subjects required to provide information. By using geographical rasters, existing statistics can also be supplemented with further characteristics (Gebers, Graze 2019: p. 12). The first official statistics with geocoding were environmental statistics and road traffic accident statistics as well as the georeferencing of agricultural holdings within the framework of the 2010 agricultural census. This was already possible before the revision of the Federal Statistics Act in 2013 (Brenner 2015: p. 48). However, since 2013, the selection of regionally structured georeferenced evaluations, which the official statistics publish free of charge, has increased considerably. By the end of 2019, more than 175 geocoded statistics were already available. A selection of these is presented in the following.

A first example of this are statistics that can be represented cartographically. These include, for example, the agricultural atlas, the census atlas, the accident atlas and the hospital atlas. The interactive agricultural atlas is based on raster maps and includes 16 agricultural maps taken from the 2010 agricultural census. For example, the average size of farms is shown in a grid of five by five kilometres (Brenzel, Gebers 2020: p. 52 f.). The census atlas is also an interactive map service with the data from the last census from 2011 on the topics of population and housing based on grids of one kilometre by one kilometre. The associated population calculator can also be used to calculate population figures for any area, such as the population figure in a noise zone around an airport (Brenzel, Gebers 2020: p. 53). The census atlas is a good example of why the grid cells are presented cartographically and not in tabular form. With grids of one kilometre in length and an area of Germany of 360,000 square kilometres, it is almost impossible to interpret the results in tables because, above all, the exact location of the grid is also not assignable for most people on the basis of the grid ID. In the cartographic representation, in addition to the statistical results, other information can also be displayed for orientation purposes, such as borders, capitals or rivers (Neutze 2015: p. 64). The accident atlas indicates the number of accidents with personal injury within the respective grids. This can help to identify particularly dangerous areas where many accidents occur (Brenzel, Gebers 2020: p. 53).

Another popular field of application of georeferenced statistics are so-called accessibility analyses. An example of this is the hospital atlas. In addition to the locations of hospitals, it also shows the transport times to the nearest hospital for any point. It is possible to differentiate between different types of hospitals, such as maternity hospitals. This makes it possible, among other things, to examine how journey times differ between urban and rural areas (Brenzel, Gebers 2020: p. 53 f.). Another example is the accessibility of primary schools. Here, the average distance of families with children of primary school age to the nearest primary school in Hesse is examined. For this purpose, the data set of the microcensus was subsequently extended by additional characteristics by means of the grid ID, which turns out to be a very low-burden method. By locating the primary schools with coordinates, the distance to the nearest primary school can be calculated for each grid cell (Gebers, Graze 2019: p. 13 ff.).

Other examples of the use of georeferencing in official statistics are, for instance, interactive population pyramids or the regional atlas, which contains regional indicators. Much of this data is collected regionally by the federal state statistical offices. However, since these data are important for the community, there is also a national spatial data infrastructure called GDI in addition to the European spatial data infrastructure INSPIRE, whereby the data can be made available to the user in a uniform manner. The "Regionaldatenbank Germany" exists as a joint project of the federal statistical office and the federal state statistical offices to provide regional data (Brenner 2015: p. 48 ff.).

Now that numerous examples have been considered in which grids are used in the context of georeferencing, the next subsection examines the advantages and disadvantages of georeferencing with grids in official statistics.

Advantages and disadvantages of using grids in official statistics

This subsection starts with the advantages of using grids in official statistics to obtain a spatial reference. A first advantage has already been mentioned in some places and relates to the reduced survey burden. Since the grid ID can be used to subsequently add characteristics to a data set, these do not have to be collected separately again (Gebers, Graze 2019: p. 12). This helps to avoid an excessive burden on respondents, as required by the ninth principle of the European Statistics Code of Practice (Eurostat 2018: p. 15 f.). Small area data are very sensitive data which, like all official statistics data, require strong protection and for which special methods of confidentiality are used in accordance with the fifth principle of the Code of Practice. The use of grids plays an important role here. They make it possible to delete the exact address and coordinates of respondents without losing the spatial reference. The deletion of the exact address is an essential point in the context of data protection (Brenzel, Gebers 2020: p. 50). Another advantage that would not be possible without the use of grids, is the borderless representation of the city and its surroundings. Through grid cells with a width of, for example, 100 metres, processes at the borders between two areas can be studied well. Moreover, the grids can be compared very accurately with each other due to their uniform size (Neutze 2015: p. 65 ff.).

In addition to these and some other advantages, the use of grids in georeferencing in official statistics also entails some disadvantages. One disadvantage is mainly related to the size of the grids. With grids of 100 by 100 metres, it can happen that there are only very few units in a grid. Through data altering confidentiality procedures, high relative deviations can occur, especially with low occupancy numbers, so that the proven values of a single cell are only reliable to a limited extent (Neutze 2015: p. 65). Another disadvantage is that natural and administrative boundaries are lost through the grid-based representation. However, since many political decisions are made at the administrative level, data in the grids are less helpful here than if they were available for the administrative areas. In this case, grid-based evaluations only serve as a supplement and cannot completely replace administrative analyses (Neutze 2015: p. 67).

Overall, it can be said that grid-based georeferencing in official statistics has both advantages and disadvantages. However, the many application examples, only some of which have been presented in this section, show the important role that georeferencing has now taken in official statistics. One aspect that also has an influence here and which is driving georeferencing forward is digitalisation and the emergence of Big Data. Therefore, georeferencing will be examined in this context in the next section.

3.4 Georeferencing in the environment of digitalisation and Big Data

Having presented explicit analyses based on georeferencing, it should be mentioned that georeferencing is an important component of the digitisation of official statistics. In the digital agenda of the Federal Statistical Office, which was published in 2019 and deals with the new framework conditions of digitalisation, one point within the framework of automated data processing is the establishment of georeferencing in all central and decentralised statistics (Statistisches Bundesamt 2019: p. 25).

In the context of digitalisation, much larger amounts of digital data are being collected. In the phenomenon known as Big Data, governments and public institutions, private companies, associations and even citizens create a huge range of digital imprints (Radermacher 2019: p. 120 ff.). In order to meet the demands of the digital agenda, more statistics also need to be georeferenced due to the large volume of data. But the crucial point is that the data comes from many new sources. Thus, many different types of data exist, which is expressed by the term variety as one of the three Vs of the Big Data concept. The internet has made a major contribution to the exponential growth of digital data. Through a variety of devices, such as smartphones, tablets or notebooks, many people are connected to the internet all day long, leaving a digital footprint all the time. But sensors have also strongly contributed to this development. For instance, urban sensors, retail scanners or public transport card readers generate a large amount of data in everyday life. One type of data that is generated in this way is location data, where mainly through the use of mobile phones the data can be assigned to the position of the user. GPS or WiFi points, for example, play a role here. Due to the assigned position, these data are already georeferenced (Blazquez, Domenech 2018: p. 101 ff.). This makes it clear that through digitisation and the new variety of data sources, georeferencing is also becoming increasingly important.

In the following, an example is presented of how mobility during the Covid pandemic can be studied using modern data sources. In order to prevent illnesses and an overload of the health system, a wide variety of infection control measures have been taken since March 2020. The focus here was particularly on reducing social contacts. One indicator that can describe the number of contacts well is the mobility of the population. In order to be able to assess the implementation and effect of this measure, small-scale information is therefore needed. A very current possibility for this is aggregated and anonymised mobile phone data, which depicts the mobility of the population (Bohnensteffen et al. 2021: p. 90). For this purpose, the Federal Statistical Office publishes mobility indicators via the experimental data section. The data comes from the Telefonica Germany network, which has a market share of around 30 per cent of the 150 million SIM cards registered in Germany. Mobile data are recorded as event data in the context of the use of mobile devices. All interactions between a mobile device and a transmission mast are recorded. Mobility is then defined by the changes between the different mobile radio cells. By specifying a mobile cell, the data contain georeferences. On the one hand, the Federal Statistical Office receives data at the district level with the daily aggregated number of movements between the start and end regions. On the other hand, it receives the hourly number of outbound, inbound and round trips in a grid. Thus, on the one hand, data is available on an administrative level as well as data on a grid level. Due to anonymisation and aggregation, the identification of individuals is not possible (Bohnensteffen et al. 2021: p. 91). Using these data deliveries, the Federal Statistical Office then calculates the mobility and compares it, for example, with the years before the pandemic. This example shows the many possible applications of the new data sources and the role that georeferencing plays in this.

One challenge that must not be forgotten with the large amounts of new data is the protection of data and the privacy of the individual (Radermacher 2019: p. 121). Here, too, the use of grids could help. As has already been shown, this contributes to data protection, as it means that no exact locations and coordinates are published. The example with mobile phone data has also shown that both aggregation at the administrative level and at the grid level ensure anonymisation of the data.

Furthermore, Radermacher (2019: p. 128) also refers to smart statistics, by which he means investing in new methods and modern algorithms to be able to guarantee the quality of data and statistics in the future. One possible method that could be considered here is georeferencing, as its potential with digital data and its large field of application has already been shown in several points in this thesis. Also Radermacher (2019: p. 124) mentions the new technical possibilities and data sources, such as geo-coded data or remote sensing, to produce local and regional statistics of a high quality.

These were just a few of the examples in which official statistics uses georeferencing today. In the context of this thesis, it is also interesting to see how deficient data affect georeferencing. This will be examined in a separate chapter 6, where georeferencing and small area estimation will be considered in the light of deficit data. But before that, the next chapter gives a detailed introduction to small area estimation.

Chapter 4

Introduction to small area estimation

One area of statistics in which georeferenced data play an important role is small area estimation. Here, estimates for small areas are to be carried out with the help of data that have a location reference. A current example where the values for small areas are of great relevance is Covid reporting. Many measures and rules do not apply to the whole of Germany but vary greatly from region to region. A crucial parameter here is the 7-day incidence at district level. At times, this has even determined whether a lockdown is imposed for a district or not. However, many other measures also depend heavily on this small-scale value, which is why it was often considered more important than the incidence at the federal or state level.

4.1 Increasing demand for small-scale estimation methods

This example of Covid and the introductory thoughts on the different regional purchase prices for land in Germany based on current data from official statistics show that the need for information at smaller regional levels is a very up-to-date topic. Therefore, this chapter on small area estimation will start with the increasing demand for regionalised information and the developments in recent years. This is a phenomenon that occurs throughout research, politics and business. Everywhere, more and more indicators and statistical parameters are needed, which are broken down more deeply regionally. A popular example where regional studies and estimates are particularly common is poverty measurement (Münnich et al. 2013: p. 151 ff.). In the United States, there is a separate survey, the Small Area Income and Poverty Estimates (SAIPE) program, which is conducted by the U.S. Census Bureau and provides poverty and income estimates, such as the total number of people in poverty or the median household income for each state, county and school district (United States Census Bureau 2022). But also in Europe, many projects and research programmes are funded on regionalised information, such as EURAREA, which deals entirely with small area statistics. In addition to poverty measurement, the census is a well-known example of official statistics that now uses small area methods. Instead of the classic full census, many countries, including Germany, now only use random samples. In order to be able to work with these samples from the individual areas, new estimation methods are needed, for instance to reduce the underestimation of certain population groups in particular areas. In all these examples and many other fields of politics and economics, an increasing need for detailed information for small-scale areas becomes visible. At the same time, however, the budget for data collection does not increase, but remains the same or even decreases. Therefore, new methods are needed to combat this problem. The methods of small area statistics, which are introduced in this chapter, are particularly helpful here (Münnich et al. 2013: p. 151 ff.). The next section begins with the basics of these methods.

4.2 The basics of small area estimation

To start with the basics of small area estimation, it is useful to take a closer look at the three words that make up the expression. This procedure is started from the back to the front. The term *estimation* is to be understood here in a statistical context. One wants to infer a characteristic value or parameter of interest from a given sample or available data to the entire population. In the context of small area estimation, the population of interest can also refer only to a sub-area.

For this, the term *area* needs to be clarified. A concept that is very similar in this context is the term domain. These two expressions represent a sub-population of the total population of interest. In a geographical context, the term area is more commonly used, whereas the word domain stands for socio-demographic groups (Münnich et al. 2013: p. 157). Examples of areas are regions, provinces, counties, districts, municipalities, school districts, health service areas or metropolitan areas. A socio-deomographic domain would be, among others, a specific age-sex-race group in a large geographical area or a collection of business firms (Rao, Molina 2015: p. 1).

Now that the term area has been clarified, it will be analysed what *small area* means in this context. Generally, there is no clear, unambiguous definition of when an area is regarded as small. However, an area is considered small if the sample size in that area is insufficient to produce direct estimators with adequate precision. Conversely, this means that at these sample sizes the direct estimators produced have too large variances. In this context, a direct estimator only uses information that is available in the area under consideration. It can even happen that the sample of an area has no element at all. In this case, no estimation is possible. Other expressions for small areas include for instance local area, subdomain or small subgroup. Based on this concept, an area is considered large enough if the sample in this domain provides direct estimators with adequate precision (Rao, Molina 2015: p. 2).

If one combines the three definitions or explanations of the individual terms, the core objective of small area estimation also becomes apparent. The aim is to obtain precise estimators even in areas or domains with a small sample size, so that valid and reliable results can also be obtained for small-scale areas. The estimators of the small area estimation must take into account that the total sample size is divided among the sub-populations. In order to better deal with small sample sizes, small area statistics use *indirect estimators* that not only use information from the area under consideration, but also information from other areas. This principle is known as *borrowing strength* (Rao, Molina 2015: p. 2). Another pair of terms often used in small area estimation are *planned areas* and *unplanned areas*. In the first case of planned areas, the sample sizes are fixed and the subsample sizes of each area are predefined. In the case of unplanned areas, there are no specifications for the sample sizes. As a result, it can happen that individual areas have no observations and are referred to as *non-sampled areas* (Münnich et al. 2013: p. 157 f.). Another distinction that must be made in the field of small area estimation and which concerns the choice of model depends on the aggregation level of the variables. If the information is only available on an aggregated level for the entire area, *area-level models* are required. If the variables of interest are available at the individual level, *unit-level models* can be used (Hobza et al. 2021: p. 3).

At the end of this section, the most important notations in the context of areas and small area estimation are given. The index d is used to identify the individual areas. Thus there is a total of $1, \dots, D$ areas. The entire population is denoted by U and U_d stands for the subpopulations in the individual areas. The total is $U = \cup_{d=1}^{D} U_d$. The samples in the individual areas are denoted by s_d and $s = \cup_{d=1}^{D} s_d$ applies. The sample size of an area is denoted by n_d and N_d refers to the population size of the area. For the total sample size $n = n_1 + \dots + n_D$ applies (Hobza et al. 2021: p. 16). Now that an introductory look at the basics of small area estimation has been given, the remainder of this chapter will analyse which means of estimation are available. First, the direct estimators mentioned above will be considered as a basis for understanding.

4.3 Short overview of direct estimators

One way to infer the whole population from survey samples and avoid expensive full surveys is to use direct estimators. If the finite population can be broken down into domains, the global estimators can be adapted for the individual domains. If the domains are treated as independent new populations, they are called direct estimators. This also means that only the data of the target variable in the domain of interest is used and its properties are investigated with regard to the probability function of the sample design. Data from other areas are ignored entirely (Hobza et al. 2021: p. 13).

For direct estimators, one is often interested in the total value $Y_d = \sum_{i \in U_d} y_i$ or the mean value $\bar{Y}_d = \frac{1}{N_d} \sum_{i \in U_d} y_i = \frac{Y_d}{N_d}$ of an area. If one uses a design-based direct estimator that takes the sample design into account, an estimator for the total value $\hat{Y}_d$ is design-unbiased if its expected value corresponds to the total value Y_d. Thus, $\mathbb{E}(\hat{Y}_d) = Y_d$ holds. An estimator is even design-consistent if for increasing sample size the bias of $\frac{\hat{Y}_d}{N_d}$ and the variance $\frac{1}{N_d^2}\mathbb{V}(\hat{Y}_d)$ tend to zero (Rao, Molina 2015: p. 10 f.).

The Horvitz-Thompson estimator

A very well-known direct estimator that is design unbiased is the Horvitz-Thompson estimator. This estimator for the total value of an area is given by:

$$\hat{Y}_d^{HT} = \sum_{i \in s_d} w_i y_i \; = \sum_{i \in s_d} \frac{1}{\pi_i} y_i.$$

The Horvitz-Thompson estimator uses the design weights w_i, where $w_i = \frac{1}{\pi_i}$ with π_i as the first-order inclusion probability that the i-th element will be in the sample s_d (Münnich et al. 2013: p. 153 ff.). The Horvitz-Thompson estimator is unbiased because its expected value is $\mathbb{E}(\hat{Y}_d^{HT}) = Y_d$. The variance of the Horvitz-Thompson estimator of a total value is:

$$\mathbb{V}(\hat{Y}_d^{HT}) = \sum_{i \in U_d} \pi_i(1-\pi_i) \cdot \left(\frac{y_i}{\pi_i}\right)^2 + 2 \sum_{\substack{i \in U_d \\ i<j}} \sum_{j \in U_d} (\pi_{ij} - \pi_i \pi_j) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j}.$$

Here, π_{ij} are the second-order inclusion probabilities that both the i-th and j-th elements are included in the sample together. The variance $\mathbb{V}(\hat{Y}_d^{HT})$ can be estimated by (Münnich et al. 2013: p. 154):

$$\hat{\mathbb{V}}(\hat{Y}_d^{HT}) = \sum_{i \in s_d} (1-\pi_i) \cdot \left(\frac{y_i}{\pi_i}\right)^2 + 2 \sum_{\substack{i \in s_d \\ i<j}} \sum_{j \in s_d} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}}\right) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j}.$$

The Horvitz-Thompson estimator for the mean is then given by (Hobza et al. 2021: p. 17):

$$\hat{\bar{Y}}_d^{HT} = \frac{\hat{Y}_d^{HT}}{N_d}.$$

The generalized regression estimator

With the Horvitz-Thompson estimator, it is sufficient if only the variable of interest is present for the observations. However, if there are also values for other characteristics, the generalised regression estimator can as well be used. The use of additional variables is central to small area estimation and is clearly visible later in the estimators presented. The generalised regression estimator, as a model-assisted estimator, uses the information from a regression model to correct the Horvitz-Thompson estimate and is given in its general form by (Münnich et al. 2013: p. 155):

$$\hat{Y}^{GREG} = \hat{Y}^{HT} + (X - \hat{X}^{HT})\hat{\beta}. \tag{4.1}$$

Here $X = (X_1, \ldots, X_p)^T$ denotes the known population totals of the auxiliary variables. Furthermore, it is assumed that for the elements of the sample the auxiliary vector x_i is also known, so that for each element of the sample $i \in s$ the data combinations (y_i, x_i) are observed. The matrix $\hat{X}^{HT}$ is defined analogously to the definition of the Horvitz-Thompson estimator as $\hat{X}^{HT} = \sum_{i \in s} w_i x_i$. The $\hat{\beta}$ vector $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ is the solution of the weighted least squares equation:

$$\hat{\beta} = \left(\sum_{i \in s} w_i x_i^T x_i \right)^{-1} \sum_{i \in s} w_i x_i^T y_i. \tag{4.2}$$

The inclusion of the design weights w_i ensures an asymptotically unbiased estimation of β within the sample design. By rearranging the terms, one obtains another representation of the generalised regression estimator:

$$\hat{Y}^{GREG} = X\hat{\beta} + \sum_{i \in s} w_i \underbrace{(y_i - x_i\hat{\beta})}_{e_i}. \tag{4.3}$$

If only one auxiliary variable x is present, the generalised regression estimator simplifies to the ratio estimator (Rao, Molina 2015: p. 13 ff.)

$$\hat{Y}_R = \frac{\hat{Y}^{HT}}{\hat{X}^{HT}} X.$$

A variance estimator in its general form is given by:

$$\hat{\mathbb{V}}(\hat{Y}) = \sum_{i<j} \sum_{i,j \in s} w_{ij}(s) b_i b_j \left(\frac{y_i}{b_i} - \frac{y_j}{b_j} \right)^2.$$

The weights $w_{ij}(s)$ provide the unbiasedness. In the case of a design with $w_i = \frac{1}{\pi_i}$, $b_i = \pi_i$ and $w_{ij}(s) = \frac{(\pi_{ij} - \pi_i \pi_j)}{(\pi_{ij} \pi_i \pi_j)}$ holds. For the variance estimate of the generalised regression estimator the estimator is given via a Taylor linearization method as (Rao, Molina 2015: p. 15):

$$\hat{\mathbb{V}}(\hat{Y}^{GREG}) = \sum_{i<j} \sum_{i,j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{(\pi_{ij} \pi_i \pi_j)} \pi_i \pi_j \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j} \right)^2 \tag{4.4}$$

$$= \sum_{i<j} \sum_{i,j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i - x_i\hat{\beta}}{\pi_i} - \frac{y_j - x_j\hat{\beta}}{\pi_j} \right)^2. \tag{4.5}$$

In the context of this thesis, however, the estimators for the entire population are of less interest than those for the sub-populations. Therefore, the generalised regression estimator for the individual $1, \dots, D$ areas will be given, as it was already the case with the Horvitz-Thompson estimator. One way to adopt the generalised regression estimator for the area totals Y_d is as follows:

$$\hat{Y}_d^{GREG} = \sum_{i \in U_d} \hat{y}_i + \sum_{i \in s_d} w_i(y_i - \hat{y}_i). \tag{4.6}$$

This formula is the representation of 4.3 broken down to the areas with the predictions $\hat{y}_i = x_i^T \hat{\beta}$. The estimator $\hat{\beta}$ is again the solution of the least squares equation as in 4.2, this time using only the elements y_{di} of the respective sample s_d in the estimation. The estimator $\hat{Y}_d^{GREG}$ has the advantage that it is approximatively unbiased even for areas with a small expected domain sample size. However, the estimator is inefficient because the auxiliary variables used are not domain specific. This means that for individuals from the total sample s that do not lie in U_d, large negative residuals occur, namely $e_{di} = -x_i\hat{\beta}$ for an element $i \in s$ that does not lie in U_d (Rao, Molina 2015: p. 17 f.).

Therefore, in the next step, a generalised regression estimator for an area total Y_d is presented, using only domain specific auxiliary information. For this, the domain totals $X_d = \sum_{i \in U_d} x_i$ must be known, where $x_{di} = x_i$ if $i \in U_d$ and $x_{di} = 0$ otherwise. Then a generalised regression estimator for Y_d following 4.1 can be represented as:

$$\hat{Y}_d^{GREG^*} = \hat{Y}_d^{HT} + (X_d - \hat{X}_d^{HT})^T \hat{\beta}_d. \tag{4.7}$$

In this case, the estimator for the $\hat{\beta}_d$-vector is the solution of the following least squares equation:

$$\hat{\beta}_d = \left(\sum_{i \in s} w_i x_{di}^T x_{di} \right)^{-1} \sum_{i \in s} w_i x_{di}^T y_{di}.$$

Unlike $\hat{Y}_d^{GREG}$, $\hat{Y}_d^{GREG^*}$ is not approximately unbiased unless the domain sample size is large. The variance estimation of $\hat{Y}_d^{GREG^*}$ is obtained by replacing e_i in 4.4 by e_{di}^*:

$$\hat{\mathbb{V}}(\hat{Y}_d^{GREG^*}) = \sum_{i<j} \sum_{i,j \in s_d} \frac{(\pi_{dij} - \pi_{di}\pi_{dj})}{\pi_{dij}} \left(\frac{y_{di} - x_{di}\hat{\beta}_d}{\pi_{di}} - \frac{y_{dj} - x_{dj}\hat{\beta}_d}{\pi_{dj}} \right)^2 .$$

Since in this case the auxiliary information is domain specific, $e_{di}^* = 0$ applies to elements of the sample s that lie outside U_d and there are no large negative residuals as in the case of $\hat{Y}_d^{GREG}$. Thus, the estimator $\hat{Y}_d^{GREG^*}$ with domain specific auxiliary information is more efficient than the estimator $\hat{Y}_d^{GREG}$ for Y_d without domain specific auxiliary information, in case the expected domain specific sample size is large (Rao, Molina 2015: p. 18 ff.).

The generalised regression estimator is the basis for many estimators in small area estimation and will appear in modified forms more frequently in this thesis. Due to their simplicity and intuition, direct estimators are used in the further course as a benchmark and comparison criterion for more complicated small area estimators in order to measure their efficiency (Hobza et al. 2021: p. 13). The generalised regression estimator may have used information outside the domain in some places to estimate the regression coefficients, but it is essentially a direct estimator (Rao, Molina 2015: p. 35). In the next section, however, indirect estimators are actually introduced.

4.4 Introduction to indirect estimators

The Horvitz-Thompson estimator and the generalised regression estimator just presented are direct estimators and only use data from the respective domain for estimation. As already mentioned, in the context of small area estimation the sample sizes are frequently small in many areas, so that the estimators often become imprecise and inefficient due to their large variances (Hobza et al. 2021: p. 41 f.). Therefore, a major goal of small area estimation is to find estimators that are precise even with small sample sizes. One solution is indirect estimators, that borrow strength from other areas by using additional information from these areas, thereby increasing the effective sample size and reducing the standard errors compared to classical methods (Rao, Molina 2015: p. 35). For the indirect small area estimators, three basic approaches are presented here: design-based, model-assisted and model-based methods. Initially, it will be started with the indirect design-based techniques.

4.4.1 Design-based indirect estimators

The design based small area approach seeks indirect estimators whose properties fit well with the sampling design distribution. Although auxiliary variables from external data sources outside the target domain are used for estimation, design-based estimators are not explicitly based on models (Hobza et al. 2021: p. 2). A selection of these estimators is presented below.

Basic synthetic estimator

The first indirect design-based estimator presented in this thesis is the basic synthetic estimator. Here, in addition to the division into areas, the population is also partitioned into a small number of strata or groups by a categorical variable as auxiliary information. A possible subdivision of the areas would be, for example, according to age-sex groups. In this case, the basic synthetic estimator is a simple and efficient estimator if the target variable has a small variance within the groups, so that the mean values of the individual groups do not differ too much between the domains. The subdivision into $g = 1, \ldots, G$ groups is denoted by $U = \cup_{g=1}^{G} U_g$. For a group g within an area d, $U_{dg} = U_d \cap U_g$ holds. s_d, s_g and s_{dg} denote the sub-samples belonging to the sample s. The auxiliary information that makes the basic synthetic estimator an indirect estimator is the domain and domain-group population sizes at the aggregate level. At the non-aggregate level, these are the sampling weights w_i and the group indicator whether i is in group g (Hobza et al. 2021: p. 42 ff.). The basic synthetic estimator includes a direct estimator for the mean of a group $\bar{Y}_g = \frac{1}{N_g}\sum_{i \in U_g} y_i$, which is given by:

$$\hat{\bar{Y}}_g^{dir} = \frac{1}{\hat{N}_g}\sum_{i \in s_g} w_i y_i, \qquad \hat{N}_g = \sum_{i \in s_g} w_i.$$

Based on this direct estimator for the group mean, the basic synthetic estimator of the area total Y_d can be expressed as:

$$\hat{Y}_d^{synth} = \sum_{g=1}^{G} N_{dg} \hat{\bar{Y}}_g^{dir}. \tag{4.8}$$

N_{dg} denotes the population size of the group g in the d-th area U_{dg}. $\hat{Y}_d^{synth}$ is indeed a biased estimator for Y_d. However, if $\bar{Y}_{dg} = \bar{Y}_g$ applies to all areas, then the synthetic estimator is approximately unbiased. This is true if the groups are homogeneous internally. Furthermore, if the selection probabilities are $\pi_{ij} = \pi_i \pi_j$ and $\pi_{ii} = \pi_i$, the variance of the synthetic estimator can be expressed as:

$$\mathbb{V}_\pi(\hat{Y}_d^{synth}) = \sum_{g=1}^{G} N_{dg}^2 \mathbb{V}_\pi(\hat{\bar{Y}}_g^{dir}) \approx \sum_{g=1}^{G} \frac{N_{dg}^2}{N_g^2} \sum_{i \in U_g} \frac{1 - \pi_i}{\pi_i}(y_i - \bar{Y}_g)^2.$$

This variance can be estimated by:

$$\hat{\mathbb{V}}_\pi(\hat{Y}_d^{synth}) = \sum_{g=1}^{G} N_{dg}^2 \hat{\mathbb{V}}_\pi(\hat{\bar{Y}}_g^{dir}) = \sum_{g=1}^{G} \frac{N_{dg}^2}{\hat{N}_g^2} \sum_{i \in s_g} w_i(w_i - 1)(y_i - \hat{\bar{Y}}_g^{dir})^2.$$

Based on the formula 4.8, the basic synthetic estimator for the mean is given as:

$$\hat{\bar{Y}}_d^{synth} = \frac{1}{N_d}\sum_{g=1}^{G} N_{dg} \hat{\bar{Y}}_g^{dir} = \frac{1}{N_d}\hat{Y}_d^{synth}.$$

Under the same conditions as for the variance of the total estimate, the variance of the basic synthetic estimator for the mean is calculated as follows:

$$\mathbb{V}_\pi(\hat{\bar{Y}}_d^{synth}) = \frac{1}{N_d^2}\sum_{g=1}^{G} N_{dg}^2 \mathbb{V}_\pi(\hat{\bar{Y}}_g^{dir}) \approx \frac{1}{N_d^2}\sum_{g=1}^{G} \frac{N_{dg}^2}{N_g^2} \sum_{i \in U_g} \frac{1 - \pi_i}{\pi_i}(y_i - \bar{Y}_g)^2.$$

The estimate for this expression is in turn given by:

$$\hat{\mathbb{V}}_\pi(\hat{\bar{Y}}_d^{synth}) = \frac{1}{N_d^2}\sum_{g=1}^{G} N_{dg}^2 \hat{\mathbb{V}}_\pi(\hat{\bar{Y}}_g^{dir}) = \frac{1}{N_d^2}\sum_{g=1}^{G} \frac{N_{dg}^2}{\hat{N}_g^2} \sum_{i \in s_g} w_i(w_i - 1)(y_i - \hat{\bar{Y}}_g^{dir})^2.$$

Using the formula for the variance, the mean squared error for the basic synthetic estimator of the area mean is given by:

$$MSE(\hat{\bar{Y}}_d^{synth}) = \mathbb{V}_\pi(\hat{Y}_d^{synth}) + \left(B_\pi[\hat{\bar{Y}}_d^{synth}]\right)^2.$$

The bias B_π of the basic synthetic estimator is given by:

$$B_\pi[\hat{\bar{Y}}_d^{synth}] = \mathbb{E}_\pi[\hat{\bar{Y}}_d^{synth}] - \frac{1}{N_d}\sum_{i \in s_d} y_i \approx \frac{1}{N_d}\sum_{g=1}^{G} N_{dg}(\bar{Y}_g - \bar{Y}_{dg})$$

and can be estimated by the plug-in estimator:

$$\hat{B}_\pi[\hat{\bar{Y}}_d^{synth}] = \frac{1}{N_d}\sum_{g=1}^{G} N_{dg}(\hat{\bar{Y}}_g^{dir} - \hat{\bar{Y}}_{dg}^{dir})$$

or especially for small area estimation problems with small subsamples s_{dg} by:

$$\hat{B}_\pi[\hat{\bar{Y}}_d^{synth}] = \hat{\bar{Y}}_d^{synth} - \hat{\bar{Y}}_d^{dir}$$

(Hobza et al. 2021: p. 42 ff.).

So one can see that the basic synthetic estimator, unlike the Horvitz-Thompson estimator for example, is typically biased. However, the variance of the basic synthetic estimator is low due to the use of auxiliary information. In the next step, the post-stratified estimator, a further indirect design-based estimator of the small area estimation, is considered.

Post-stratified estimator

For this purpose, a population is again considered that is subdivided into areas and groups, as was already the case with the basic synthetic estimator. This time, however, the target variable y has a large variance within groups and the domain-group means may differ significantly, so the basic synthetic estimator would not be a good choice. The post-stratified estimator of the area total is then given by:

$$\hat{Y}_d^{pst} = \sum_{g=1}^{G} N_{dg}\hat{\bar{Y}}_{dg}^{dir}$$

with

$$\hat{\bar{Y}}_{dg}^{dir} = \frac{1}{\hat{N}_{dg}}\sum_{i \in s_{dg}} w_i y_i, \qquad \hat{N}_{dg} = \sum_{i \in s_{dg}} w_i$$

as the Hájek-type direct estimator of $\bar{Y}_{dg} = \frac{1}{N_{dg}}\sum_{i \in U_{dg}} y_i$. The post-stratified estimator is approximately unbiased. Again, under the conditions $\pi_{ij} = \pi_i\pi_j$ and $\pi_{ii} = \pi_i$, the variance of the post-stratified estimator and the corresponding estimate of it are given by:

$$\mathbb{V}_\pi(\hat{Y}_d^{pst}) = \sum_{g=1}^{G} N_{dg}^2 \mathbb{V}_\pi(\hat{\bar{Y}}_{dg}^{dir}) \approx \sum_{g=1}^{G} \sum_{i \in U_{dg}} \frac{1-\pi_i}{\pi_i}(y_i - \bar{Y}_{dg})^2$$

$$\hat{\mathbb{V}}_\pi(\hat{Y}_d^{pst}) = \sum_{g=1}^{G} N_{dg}^2 \hat{\mathbb{V}}_\pi(\hat{\bar{Y}}_{dg}^{dir}) = \sum_{g=1}^{G} \frac{N_{dg}^2}{\hat{N}_{dg}^2} \sum_{i \in s_{dg}} w_i(w_i - 1)\left(y_i - \hat{\bar{Y}}_{dg}^{dir}\right)^2$$

(Hobza et al. 2021: p. 45 f.).

These formulas can also be applied to the mean value of an area. Thus, the post-stratified estimator for the mean value results in:

$$\hat{\bar{Y}}_d^{pst} = \frac{1}{N_d} \sum_{g=1}^{G} N_{dg} \hat{\bar{Y}}_{dg}^{dir} = \frac{1}{N_d} \hat{Y}_d^{pst}.$$

Like the post-stratified estimator of area totals, this estimator for the mean is also approximately unbiased. The expressions of the variance are given by:

$$\mathbb{V}_\pi(\hat{\bar{Y}}_d^{pst}) = \frac{1}{N_d^2} \sum_{g=1}^{G} N_{dg}^2 \mathbb{V}_\pi(\hat{\bar{Y}}_{dg}^{dir}) \approx \frac{1}{N_d^2} \sum_{g=1}^{G} \sum_{i \in U_{dg}} \frac{1-\pi_i}{\pi_i}(y_i - \bar{Y}_{dg})^2$$

$$\hat{\mathbb{V}}_\pi(\hat{\bar{Y}}_d^{pst}) = \frac{1}{N_d^2} \sum_{g=1}^{G} N_{dg}^2 \hat{\mathbb{V}}_\pi(\hat{\bar{Y}}_{dg}^{dir}) = \frac{1}{N_d^2} \sum_{g=1}^{G} \frac{N_{dg}^2}{\hat{N}_{dg}^2} \sum_{i \in s_{dg}} w_i(w_i - 1)\left(y_i - \hat{\bar{Y}}_{dg}^{dir}\right)^2.$$

Due to the approximate unbiasedness of the post-stratified estimator, the mean sqaured error can be approximated via the variance:

$$MSE(\hat{\bar{Y}}_d^{pst}) \approx \mathbb{V}_\pi(\hat{\bar{Y}}_d^{pst}), \qquad mse(\hat{\bar{Y}}_d^{pst}) = \hat{\mathbb{V}}(\hat{\bar{Y}}_d^{pst}).$$

Here $mse(\cdot)$ is un upper biased estimator for the mean squared error $MSE(\cdot)$.

In summary, it can be stated that for the calculation of the post-stratified estimator, the sampling weights w_i and the domain-group indicator whether i lies in U_{dg} as well as the domain and domain-group population sizes are used as auxiliary information. In contrast to the basic synthetic estimator, the post-stratified estimator is normally unbiased, but has a larger variance (Hobza et al. 2021: p. 46 f.). Next, a brief presentation of an indirect design-based estimator is given, which combines the two estimators just considered.

Sample size dependent estimator

One way to account for both the synthetic and post-stratified components is the sample size dependent estimator. Here, both components are weighted by γ according to the population size of the areas:

$$\hat{Y}_d^{ssd} = \gamma_d \hat{Y}_d^{pst} + (1 - \gamma_d)\hat{Y}_d^{synth},$$

with

$$\gamma_d = \begin{cases} 1 & \text{if } \hat{N}_d \geq \delta N_d \\ \frac{\hat{N}_d}{\delta N_d} & \text{otherwise.} \end{cases}$$

The influence of the synthetic estimator can be controlled via the constant δ. Possible values here are, for instance, $\delta = 1$ or $\delta = 2/3$ (Hobza et al. 2021: p. 47). The sample size dependent estimator tries to combine the advantages of the synthetic and the post-stratified estimator. These two estimators have in common that they only use population sizes and indicators as well as sampling weights, but no other variables as auxiliary information. The case with the help of further variables will be looked at in the next sections.

4.4.2 Model-assisted indirect estimators

If, in addition to the target variable, other data are available as auxiliary information related to the target variable, it is possible to obtain more accurate domain estimators by using explicit models with these auxiliary variables instead of design-based methods. Model-assisted approaches are a first step in this direction. Here, the properties under the design-based distribution are still taken into account, but explicit models are also included in the choice of the estimation procedure (Hobza et al. 2021: p. 2). One model-assisted estimator has already been investigated in this thesis, namely the generalised regression estimator. Here it is often discussed whether this estimator is a direct estimator or is already considered an indirect estimator. The domain specific estimator $\hat{Y}_d^{GREG}$ from the equation 4.6 also uses auxiliary information from other areas when estimating $\hat{\beta}$, whereas the estimator $\hat{Y}_d^{GREG*}$ from the equation 4.7 uses only the auxiliary information from the corresponding area when estimating $\hat{\beta}_d$. According to Rao and Molina (2015: p. 35), the estimator was classified as a direct estimator, whereas in Hobza et al. (2021: p. 47 ff.), for example, it is already considered an indirect estimator. No matter in which group one classifies this estimator, there are plenty of other model-assisted indirect estimators.

Calibration estimators

The first type of indirect model-assisted estimators presented here are calibration estimators. The calibration estimators are a family of estimators that rely on a common base of auxiliary information. They use calibrated weights that, given a distance measure, approximate as closely as possible the original weights of the sampling frame π_i^{-1} while obeying a set of constraints, the calibration equations. For the i-th element of the sample s, (y_i, x_i) is observed, where x_i is the vector of auxiliary variables of the i-th element. Furthermore, the total value of X in the population is known. In the context of calibration, for the estimation of the population total Y, instead of the basic sampling design weights $w_i = \frac{1}{\pi_i}$ of the Horvitz-Thompson estimator, new weights w_i^* should be used for the estimator $\hat{Y} = \sum_{i \in s} w_i^* y_i$, which are as close as possible to w_i for a given metric, but also satisfy the calibration equation

$$\sum_{i \in s} w_i^* x_i = X. \tag{4.9}$$

One distance measure for the average distance between the weights w_i and w_i^* with the known positive weights $\frac{1}{q_i}$, is in the framework of the chi-square statistic

$$\mathbb{E}_\pi \left(\sum_{i \in s} \frac{(w_i^* - w_i)^2}{w_i q_i} \right). \tag{4.10}$$

The q_i must be determined in advance. A common form of this is uniform weighting $\frac{1}{q_i} = 1$ or unequal weighting $q_i = \frac{1}{x_i}$ in the case that the vector of auxiliary information consists of only one variable. Since the distance between the new weights and the sampling weights should be as small as possible, the expression 4.10 should be minimized, while the equation 4.9 should still hold for each sample s. Thus, the conditional value of the distance is to be minimized for a realised sample s, whereby the auxiliary information is included via the calibration equation. The original sample weights w_i produce unbiased estimators. Therefore, one hopes for an almost unbiased estimator by a small distance to them (Särndal, Deville 1992: p. 376 f.). By minimising the distance measure in the context of the calibration equation, one obtains the calibrated weight

$$w_i^* = w_i(1 + q_i x_i^{'} \lambda) \tag{4.11}$$

where λ as a vector of the Lagrange multiplier is determined by

$$\lambda = T_s^{-1}(X - \hat{X}^{HT}) = T_s^{-1} \left(X - \left(\sum_{i \in s} w_i x_i \right) \right)$$

with T_s^{-1} as the existing inverse of

$$T_s = \sum_{i \in s} w_i q_i x_i x_i^{'}.$$

This gives the calibration estimator of the total of the variable of interest y:

$$\hat{Y}^{cal} = \sum_{i \in s} w_i^* y_i = \hat{Y}^{HT} + (X - \hat{X}^{HT})^{'} \hat{\beta}_s \tag{4.12}$$

with

$$\hat{\beta}_s = T_s^{-1} \sum_{i \in s} w_i q_i x_i y_i$$

(Särndal, Deville 1992: p. 376 f.). Due to the second representation of the calibration estimator in the formula 4.12, this estimator can in a way also be regarded as a form of the generalised regression estimator. As a distance measure for calculating the distance of w_i and w_i^*, there are many alternatives to the formula 4.10 in distance measures $G(w, w^*)$, for instance Hellinger distance or minimum entropy distance, all of which must fulfil certain conditions, such as non-negativity or differentiability. However, the presentation of these distance measures would go beyond the scope of this subsection. Instead, the variance of the calibration estimator is considered. The variance of the calibration estimator is close to that of the generalised regression estimator. An asymptotic form of it is given by:

$$\mathbb{V}(\hat{Y}^{cal}) = \sum_{i<j} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j)(w_i E_i)(w_j E_j) = \sum_{i<j} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{E_i}{\pi_i} \right) \left(\frac{E_j}{\pi_j} \right)$$

with $E_i = y_i - x_i \beta$. This variance can be estimated by (Särndal, Deville 1992: p. 379 f.):

$$\hat{\mathbb{V}}(\hat{Y}^{cal}) = \sum_{i<j} \sum_{i,j \in s} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) (w_i^* e_i)(w_j^* e_j)$$

with $e_i = y_i - x_i \hat{\beta}$.

Now that the calibration estimator has been introduced in its basic features, it will be broken down to individual domains for small area estimation. One possibility would be to break down all formulas from this subsection of the calibration estimator straight forward to the sample of area s_d. Then the area-specific calibration estimator would result from the formula 4.12 to (Lehtonen, Veijanen 2016: 157 f.):

$$\hat{Y}_d^{cal} = \sum_{i \in s_d} w_{di}^* y_i = \hat{Y}_d^{HT} + (X_d - \hat{X}_d^{HT})' \hat{\beta}_{s_d}$$

with

$$\hat{\beta}_{s_d} = T_{s_d}^{-1} \sum_{i \in s_d} w_{di} q_{di} x_i y_i.$$

However, this estimator is not an indirect estimator because it only uses observations of the y-variable from the d-th domain. Instead, in the context of indirect estimators, a semi-indirect calibration estimator will now be used that borrows strength from other domains to reduce the mean squared error. One possibility here is to include the neighbours of the area of interest in the estimation. Neighbourhoods are all areas that have a common boundary with the specified area. Now designate $C_d \supset U_d$ a superset of the domain of interest and the corresponding sample $r_d = C_d \cap s$ is the intersection of the superset C_d with the entire sample s (Lehtonen, Veijanen 2016: 158). Then the calibration estimator of the total Y of an area is given:

$$\hat{Y}_{d,r}^{cal} = \sum_{i \in r_d} w_{di}^* y_i.$$

The corresponding calibration equation is given by:

$$\sum_{i \in r_d} w^*_{di} x_i = \sum_{i \in U_d} x_i = X_d.$$

The weights w^*_{di} minimize the following expression with I_{di} as indicator function $I\{i \in s_d\}$:

$$\sum_{i \in r_d} \frac{(w^*_{di} - I_{di} w_i)^2}{w_i}$$

and result in

$$w^*_{di} = I_{di} w_i + \lambda_d^{'} w_i x_i$$

with

$$\lambda_d^{'} = \left(\sum_{i \in U_d} x_i - \sum_{i \in r_d} I_{di} w_i x_i \right)^{'} \left(\sum_{i \in r_d} w_i x_i x_i^{'} \right)^{-1}.$$

While the variance for the area-specific generalised regression estimator can be determined analytically, this possibility no longer exists for area-specific model calibration estimators, so that bootstrap is recommended (Lehtonen, Veijanen 2016: 158 f.).

Logistic generalised regression estimator

A further model-assisted indirect estimator presented here is the logistic generalised regression estimator. The previous estimators were mostly designed for metric target variables. However, it is also common in areas that the variable of interest is categorical or binary. For example, in the context of poverty measurement, the binary indicator of whether a person is unemployed or not could be of interest to estimate the proportion of unemployed persons. The logistic generalised regression estimator helps here. In the context of small area estimation, Lehtonen and Veijanen (2016: p. 155) use a mixed logistic model with area-specific random effects $u_d \sim N(0, \sigma_u^2)$ to account for possible differences between areas. A more detailed overview of mixed models is given in the next chapter 5 with the model-based indirect estimators. With a binary target variable y, a mixed logistic model is given by:

$$\mathbb{E}(y_i \mid u_d) = P(y_i = 1 \mid u_d, \beta) = \frac{exp(x_i^{'}\beta + u_d)}{1 + exp(x_i^{'}\beta + u_d)}$$

with the vector of auxiliary variables x and β as the vector of fixed effects for all areas. Using the data, the parameters β and σ_u^2 are first estimated so that estimators $\hat{u}_d$ of the mixed effects can be calculated for all areas so that the predictions $\hat{y}_i = P(y_i = 1 \mid \hat{u}_d, \hat{\beta})$ can be made for the units in the different domains. Based on the formula for the generalised regression estimator 4.6, the mixed logistic generalised regression estimator can also be given. The latter estimates the frequency f_d of an attribute C of a categorical variable in each area. For the model formula, in addition to the design weights, the indicators $v_i = I\{y_i \in C\}$, indicating which attribute of the categorical variable a unit takes on. With the fitted values $\hat{p}_i = P(v_i = 1 \mid \hat{u}_d; x_i, \hat{\beta})$ the mixed logistic generalised regression estimator of the class frequencies in U_d is obtained by:

$$\hat{f}_{d;MLGREG} = \sum_{i \in U_d} \hat{p}_i + \sum_{i \in s_d} w_i(v_i - \hat{p}_i)$$

The calculation of the individual $\hat{p}_i$ requires auxiliary variables x that are available at unit level (Lehtonen, Veijanen 2016: 155 f.). The calibration estimator and the mixed logistic generalised regression estimator are examples of model-assisted estimators. Even though they are not design-unbiased, they partly show a considerably lower variance than the desgin-based estimators. In the next step, estimators are considered that no longer take the design-based distribution into account, but are only based on models.

4.4.3 Model-based indirect estimators

Besides the design-based and the model-assisted approach, there is another possibility to obtain estimators for small-scale areas, namely via model-based methods. Here it is assumed that the data is generated by a true model, so inference should be based on it. In model-based methods, cross-sectional or temporal auxiliary information as well as spatial correlation can also be taken into account. With this approach, however, a lot depends on the quality of the model. The properties of the estimator are very good, related to the true model distribution. This means that a model that fits the data well, guarantees good model-based estimators whose mean squared errors are lower than those of the design-based estimators (Hobza et al. 2021: p. 3). Within the framework of small area estimation, there are two central estimators whose main difference is the level at which the information is available. If the required information is available at the individual level, the Battese-Harter-Fuller estimator can be used as a unit-level model. The Fay-Herriot estimator, on the other hand, uses only aggregate-level information as an area-level model (Münnich et al. 2013: p. 163). These two estimators are so central to small area estimation that the next chapter is devoted to them, as they are too comprehensive for a subsection. Instead, this introductory chapter to small area estimation will be concluded with a few application examples from official statistics.

4.5 Small area estimation in official statistics

After some examples of small area estimation from politics, economics and research have already been given at the beginning of this chapter, this chapter will look at other applications in which small area methods are explicitly used in official statistics, in addition to the examples of poverty measurement and the census already mentioned.

The first example presented here is similar to the introductory idea and aims to provide estimates of average existing rents at the municipal level in North Rhine-Westphalia using small area methods. Since there has been an emotional discussion about the development of rents in Germany for years, it is necessary to have valid information for this even at the smallest level. However, the data that the official statistics have published so far in the context of the microcensus in North Rhine-Westphalia are only available at the district level, as the sample sizes are too small in most municipalities. With the help of geodata from the microcensus, however, it is to be tested how the official statistics can also offer values at the municipality level in the future through the use of small area methods. For this purpose, a Fay-Herriot model is used, which uses characteristics from the state database that are available at the municipality level, such as population structure, election results or socio-economic status, as auxiliary information. This made it possible to generate good quality estimates for rent at the municipality level, and values could also be estimated for municipalities without sample information from the microcensus (Landesbetrieb Information und Technik Nordrhein-Westfalen 2023).

In general, the estimation of values at the level of cities and functional urban areas is a frequent objective of official statistics in Europe, although most surveys are planned in such a way that they only allow for a reliable design-based estimate at the state or federal state level. Therefore, to improve the quality of estimators at small levels, special methods are needed to obtain reliable estimates even for cities with a small sample size or even non-sampled areas at all. Thus, Eurostat has published as assistance to the national statistical authorities guidelines on small area estimation for city statistics and other functional geographies (Münnich et al. 2019: p. 7 ff.). These guidelines also contain a practical example of official statistics, where the share of persons at risk of poverty or social exclusion in 1,592 municipalities is to be estimated. The municipalities are not taken into account in the sampling design, so that these are unplanned areas. Area-level variables are used as auxiliary information, such as the share of native-born persons, the share of unemployed persons and other socio-economic characteristics. Based on this information, the Fay-Harriot model is used to estimate the value of the share of persons at risk of poverty or social exclusion in the individual cities and municipalities (Münnich et al. 2019: p. 24 ff.).

A third example of the use of small area estimation in official statistics is the regional evaluation of business statistics. This is associated with particular difficulties. Business data are often dominated by single larger outliers from a skewed distribution. Business registers are a useful source of auxiliary variables when using small area methods in the context of business statistics. In practice, however, major inconsistencies often exist between the register and the target population (Manecke 2019: p. 144). In the context of business statistics, one encounters not only a regional stratification but also an business-specific stratification, as it has already appeared in some other places in this work. The population is broken down into sub-populations according to industry groups or size classes. For all these sub-populations, the total value of the area is to be estimated. Since the auxiliary information is available at unit level, the Battese-Harter-Fuller estimator is used. Thus, despite the existing difficulties, a more stable estimation can be achieved than with design-based approaches (Manecke 2019: p. 147 ff.).

These are only three selected examples that show how important small area methods have become in official statistics. In all three of these examples, either the Fay-Herriot or Battese-Harter-Fuller estimator is used, which illustrates the central role they play in small area estimation. Therefore, the next chapter is devoted to them.

Chapter 5

The two major estimators of small area estimation

This chapter is mainly focused on the Battese-Harter-Fuller estimator and the Fay-Herriot estimator, which have already appeared in some parts of this thesis. Like the mixed logistic generalised regression estimator, these two estimators use as model-based estimators mixed models in area estimation. Therefore, this chapter begins with a brief overview of mixed models as a further foundation.

5.1 Basics of mixed models

Mixed models always become important when data are available in longitudinal or clustered form. The different areas of small area estimation can be considered as clusters. In such cases, the models should contain subject- or cluster-specific effects, which are also referred to as random effects, in addition to the fixed population effects β (Fahrmeir et al. 2013: p. 38 ff.). The data of the sample elements in the $d = 1, \dots, D$ areas with the respective sample sizes n_d result for the target variable y and the auxiliary variables $x_1, \dots, x_k$ to:

$$(y_{di}, x_{di1}, \dots, x_{dik}), \qquad i = 1, \dots n_d.$$

Thus, the mixed model is given by:

$$y_{di} = \beta_0 + \beta_1 x_{di1} + \cdots + \beta_k x_{dik} + z_{0d} + z_{1d} u_{di1} + \cdots + z_{qd} u_{diq} + \epsilon_{di} \tag{5.1}$$

or in matrix notation:

$$y = x\beta + zu + \epsilon.$$

Here y, x and β are defined as in the classical linear regression model. The matrix z is assumed to be known and the random effects are represented by the vector u. In the context of small area estimation, z is often used as an indicator that assigns the elements to a certain area. The assumption is made that both the disturbance term ϵ and u are each multivariate normally distributed. Both ϵ and u are expected to have an expectation of zero and the variance-covariance matrices are given by Σ_ϵ, respectively Σ_u. Furthermore, ϵ and u are supposed to be stochastically independent of each other. The maximum likelihood estimator for β is then given by $\hat{\beta} = (x^{'}v^{-1}x)^{-1}x^{'}v^{-1}y$, where $v = \Sigma_\epsilon + z\Sigma_u z^{'}$ (Münnich et al. 2013: p. 163 f.). Observations in different clusters or areas are still considered independent, but a dependency is assumed between units from the same area as they share common properties. This gives rise to two sources of variation, namely between and within area variation, which makes flexible modelling of the data possible. The mixed models thus take into account the spatial dependence of the areas and the area random effects explain the variance between the areas, which cannot yet be expressed by the fixed effects. The use of random effects in the mixed model also has the advantage that only one variance parameter has to be estimated for each area instead of a separate intercept (Hobza et al. 2021: p. 4). The models differ externally depending on the level at which the variables are present. For example, the representation 5.1 is only possible if the covariates x are present at unit-level. This is also taken into account by the Battese-Harter-Fuller estimator and the Fay-Herriot estimator, which are now examined after the basics of mixed models. It will be started with the Battese-Harter-Fuller estimator and the case where the auxiliary variables are present at unit-level.

5.2 Battese-Harter-Fuller (BHF) estimator for unit-level models

The basic unit-level estimator of small area estimation is the Battese-Harter-Fuller estimator with its underlying nested error regression model. It is required that the auxiliary variables $x_{di} = (x_{di1}, \ldots, x_{dik})$ are available for each population element i of each small area d, although it is often sufficient if only the population means $\bar{X}_d$ are available. The target variable y_{di} is related to x_{di} via the nested error linear regression model:

$$y_{di} = x_{di}^T\beta + u_d + \epsilon_{di}.$$

It holds that the area-specific effects u_d are independent and identically distributed random variables with $\mathbb{E}(u_d) = 0$ and $\mathbb{V}(u_d) = \sigma_u^2$. The noise terms ϵ_{di} must satisfy the condition $\epsilon_{di} = k_{di}\tilde{\epsilon}_{di}$ for known constants k_{di} with $\tilde{\epsilon}_{di}$ as independent and identically distributed random variables, which are independent of u_d and for which $\mathbb{E}(\tilde{\epsilon}_{di}) = 0$ and $\mathbb{V}(\tilde{\epsilon}_{di}) = \sigma_\epsilon^2$ hold (Rao, Molina 2015: p. 78 f.). Under these conditions, the BHF estimator for estimating the area means is given by:

$$\hat{\bar{Y}}_d^{BHF} = \bar{X}_d\hat{\beta} + \hat{u}_d. \tag{5.2}$$

Battese, Harter and Fuller were the first to use the nested error regression model for small area problems to predict corn and soybean areas for twelve counties in north-central Iowa using satellite data. Therefore, the estimator 5.2 is named after them and the nested error regression model is considered the basic unit-level model in small area estimation (Hobza et al. 2021: p. 155). Within the framework of this model, the estimator for β is again $\hat{\beta} = (x^{'}v^{-1}x)^{-1}x^{'}v^{-1}y$ and the estimated random effects are:

$$\hat{u}_d = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\epsilon^2}{n_d}}(\bar{y}_d - \bar{x}_d\hat{\beta}). \tag{5.3}$$

The vector $\bar{X}_d$ represents the population average value of auxiliary characteristics in each area d and $\bar{x}_d$ is the unweighted mean value of the auxiliary information in the sample of the d-th area. If one now substitutes 5.3 into 5.2, one obtains the following form of the BHF estimator:

$$\hat{\bar{Y}}_d^{BHF} = \gamma_d(\bar{y}_d + (\bar{X}_d - \bar{x}_d)\hat{\beta}) + (1 - \gamma_d)\bar{X}_d\hat{\beta} \tag{5.4}$$

with

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\epsilon^2}{n_d}}.$$

The form 5.4 shows that the BHF is a composite estimator. The first part is a direct estimator as a multilevel generalised regression estimator, whereas the second part of 5.4 is a synthetic estimator. The previous chapter already showed that the synthetic estimator is typically biased, which is why the BHF estimator usually has a bias. However, the BHF estimator has the property of being the best linear unbiased predictor (BLUP). The more precise meaning of best predictors will be discussed in more detail later in this chapter in 5.5. At this point, it should only be noted that the BHF is an unbiased predictor and has the smallest variance of all linear unbiased predictors (Münnich et al. 2013: p. 164 f.). The exact variance of the BHF estimator is usually unknown, since the individual variance components are unknown in practice and must first be estimated. In practice, this variance estimation is automatically taken into account when estimating the mixed model with software packages. In theory, however, this is very technical and is relegated to the section of the best linear unbiased estimators 5.5. The estimates $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_u^2$ replace σ_ϵ^2 and σ_u^2, making γ_d to $\hat{\gamma}_d$ and the BHF becomes the empirical best linear unbiased predictor (EBLUP). In the BHF model, the sample information is required at the unit level, i.e. at the smallest disaggregated unit. For the register information, on the other hand, it is sufficient to have it available in aggregated form at area level due to the linearity of the estimator (Münnich et al. 2013: p. 165). An estimator that only requires aggregate data for the entire model estimation is the Fay-Herriot estimator and this is introduced in the next section.

5.3 Fay-Herriot (FH) estimator for area-level models

For the basic area model it is considered that $\theta_d = g(\bar{Y}_d)$, for some specified $g(\cdot)$, is connected to the area specific auxiliary data $x_d = (x_{d1}, \ldots, x_{dk})$ via the following linear model:

$$\theta_d = x_d^T \beta + z_d u_d, \qquad d = 1, \ldots, D.$$

Again, the area-specific effects u_d are assumed to be independent and identically distributed with $\mathbb{E}(u_d) = 0$ and $\mathbb{V}(u_d) = \sigma_u^2$, so $u_d \overset{iid}{\sim} (0, \sigma_u^2)$. It is possible to consider the parameter σ_u^2 as a measure of the homogeneity of the individual areas after accounting for the auxiliary variables x_d (Rao, Molina 2015: p. 76). In order to be able to perform inference on the small area means $\bar{Y}_d$ under these model conditions, it is assumed that direct estimators $\hat{\bar{Y}}_d$ are available, from which it follows:

$$\hat{\theta}_d = g(\hat{\bar{Y}}_d) = \theta_d + \epsilon_d, \qquad d = 1, \ldots, D.$$

Again, the sampling errors ϵ_d are independent with $\mathbb{E}(\epsilon_d \mid \theta_d) = 0$ and $\mathbb{V}(\epsilon_d \mid \theta_d) = \sigma_\epsilon^2$. If one now combines the previous formulas and results from this section, one obtains the basic area level model:

$$\hat{\theta}_d = x_d^T \beta + z_d u_d + \epsilon_d. \tag{5.5}$$

Here the u_d and ϵ_d are independent of each other and the model 5.5 is a special case of the linear mixed model (Rao, Molina 2015: p. 76 f.). The first users of this model in small area estimation were Fay and Herriot (1979: p. 272 ff.). They use the model 5.5 to estimate the log per capita income for small places in the United States with less than 1,000 inhabitants. For their estimation they set $z_d = 1$ and the following estimator is named after them:

$$\hat{\bar{Y}}_d^{FH} = \bar{X}_d \hat{\beta} + \hat{u}_d. \tag{5.6}$$

The data for the Fay-Herriot estimator is usually known at area level as area averages. These are used to estimate $\hat{\beta}$. With the area means, an estimator for the random effects can also be reproduced:

$$\hat{u}_d = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\epsilon^2}{n_d}} (\bar{y}_d - \bar{X}_d \hat{\beta}). \tag{5.7}$$

The only difference to the random effects estimator for the BHF 5.3, besides the estimation of β, is the use of the population mean of the auxiliary data $\bar{X}_d$ in 5.7 instead of the sample mean $\bar{x}_d$ in 5.3. As with the BHF, the FH can be represented as a composite estimator by substituting 5.7 in 5.6:

$$\hat{\bar{Y}}_d^{FH} = \gamma_d \bar{y}_d + (1 - \gamma_d) \bar{X}_d \hat{\beta}. \tag{5.8}$$

As with the composite BHF estimator, γ again corresponds to $\gamma = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\epsilon^2}{n_d}}$. Using the estimated variance components $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_u^2$ again, γ_d becomes $\hat{\gamma}_d$. Thus, the Fay-Herriot estimator also becomes the empirical best linear unbiased predictor (Münnich et al. 2013: p. 165 f.). The difference to the BHF estimator is that the composite FH estimator in the first part does not take into account the difference between population mean and sample mean in the areas multiplied by $\hat{\beta}$. Thus, the first part is no longer a multilevel generalised regression estimator but a Horvitz-Thompson estimator. The second part of the composite FH estimator is still a synthetic estimator.

5.4 Comparison of the two estimators

Now that both estimators have been fundamentally introduced, they will be compared with each other before the theory of the two estimators is explored in more depth. As has been pointed out in many parts of this thesis, the major difference between the two estimators is the level of information at which the data are available. Thus, the advantages and disadvantages of the two estimators in direct comparison with each other mainly relate to this point. A major advantage of using area level data in the Fay-Herriot estimator is that it requires relatively little information. This is also a benefit in the light of data protection. In the second chapter on deficient data, it was elaborated how strongly microdata must be protected in order to follow data protection principles, especially in official statistics. Since the FH estimator only requires aggregated data, the protection of microdata does not need to be taken into account when using the FH estimator. However, the use of aggregated data can also be a disadvantage of the FH estimator compared to the BHF estimator. The aggregation of data is usually accompanied by a considerable loss of information, which means that in many cases the FH estimator provides worse estimates than the BHF estimator. But, nowadays, in times of Big Data, more and more information is available from different sources. At unit level, this information must first be linked to each other in a complex process using matching methods and linkage procedures. At area level, on the other hand, aggregated information from several sources can easily be used in the FH estimator, which makes it easy to create extensive models. In this way, the loss of information just described could at least be compensated for in comparison to the BHF estimator (Münnich et al. 2013: p. 166).

5.5 (Empirical) best linear unbiased prediction for small area estimation

Now that both the Battese-Harter-Fuller and the Fay-Herriot estimator have been introduced in their basic features and compared with each other, this section considers further properties of the two estimators and also focuses on particularly desirable characteristics. Both estimators were stated to be EBLUP. Since both are a special case of a mixed linear model, it is first considered what empirical best linear unbiased predictor means for models with fixed and random effects. It is started with the nested error regression model, which serves as the basis for the BHF estimator.

Best linear unbiased predictors under the nested error regression model

At the beginning of the section on the Battese-Harter Fuller estimator, the nested error regression model at unit level, was given by $y_{di} = x_{di}^T\beta + u_d + \epsilon_{di}$ for $d = 1, \dots, D$ and $i = 1, \dots, N_d$. Going one step further on domain level, the model can be expressed by $y_d = X_d\beta + Z_d u_d + \epsilon_d$ with $Z_d = 1_{n_d}$, $u_d \sim N(0, \sigma_u^2)$ and $\epsilon_d = \text{col}(\epsilon_{di}) \sim N(0_{n_d}, \sigma_\epsilon^2 W_d^{-1})$, where the W_d are known heteroscedasticity weights. The variance matrix for the y_d in the areas is then given by:

$$V_d = var(y_d) = \sigma_u^2 1_{n_d} 1_{n_d}^{'} + \sigma_\epsilon^2 W_d^{-1}.$$

Then it is possible to move on to the nested error regression model in matrix form:

$$y = X\beta + Zu + \epsilon,$$

where $y = \text{col}(y_d)$, $X = \text{col}(X_d)$, $Z = \text{diag}(Z_d)$, $V_u = \sigma_u^2 I_D$, $V_\epsilon = \sigma_\epsilon^2 W^{-1}$ and $W = \text{diag}(W_d)$. The variance matrix for y is then composed of:

$$V = var(y) = ZV_uZ^{'} + V_\epsilon = Z(\sigma_u^2 I_D)Z^{'} + \sigma_\epsilon^2 W^{-1}.$$

When calculating BLUPs, the inverse of the variance is also needed, which is given by:

$$V_d^{-1} = \frac{1}{\sigma_\epsilon^2}\left(W_d - \frac{\gamma_d^w}{w_d} w_{n_d} w_{n_d}^{'}\right),$$

with $\gamma_d^w = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\epsilon^2}{w_d}}$ (Hobza et al. 2021: p. 156 f.). Now, assuming that the variance components σ_ϵ^2 and σ_u^2 are positive, the best linear unbiased estimator (BLUE) for β is given by

$$\tilde{\beta} = (X^{'}V^{-1}X)^{-1}X^{'}V^{-1}y$$

and the best linear unbiased predictor (BLUP) of the random effects is

$$\tilde{u} = V_u Z^{'} V^{-1}(y - X\tilde{\beta}).$$

A more efficient calculation for the BLUE of β is as a sum over the areas:

$$\tilde{\beta} = \left(\sum_{d=1}^{D} X_d^{'} V_d^{-1} X_d\right)^{-1} \left(\sum_{d=1}^{D} X_d^{'} V_d^{-1} y_d\right).$$

A simpler and intuitive formula for the BLUP of the random effects is given by:

$$\tilde{u}_d = \gamma_d^w \left(\hat{\bar{Y}}_d^w - \hat{\bar{X}}_d^w \tilde{\beta} \right),$$

where $\hat{\bar{Y}}_d^w = \frac{1}{w_d} \sum_{i=1}^{n_d} w_{di} y_{di}$ and $\hat{\bar{X}}_d^w = \frac{1}{w_d} \sum_{i=1}^{n_d} w_{di} x_{di}$.

Thus, the BHF estimator as BLUP is analogous to the formula 5.2 given by $\tilde{\bar{Y}}_d^{BHF} = \bar{X}_d \tilde{\beta} + \tilde{u}_d$. Substituting the BLUE for β and the BLUP for u leads to the following expression:

$$\begin{aligned}
\tilde{\bar{Y}}_d^{BHF} &= \bar{X}_d (X^{'} V^{-1} X)^{-1} X^{'} V^{-1} y + V_u Z^{'} V^{-1} (y - X\tilde{\beta}) \\
&= \bar{X}_d \left(\sum_{d=1}^{D} X_d^{'} V_d^{-1} X_d \right)^{-1} \left(\sum_{d=1}^{D} X_d^{'} V_d^{-1} y_d \right) + \gamma_d^w \left(\hat{\bar{Y}}_d^w - \hat{\bar{X}}_d^w \tilde{\beta} \right).
\end{aligned}$$

In practice, the variance components σ_u^2 and σ_ϵ^2 are unknown and thus the BLUPs are not calculable under linear models with random effects. To address this problem, the variance components are consistently estimated by $\hat{\sigma}_u^2$ and $\hat{\sigma}_\epsilon^2$, giving the empirical BLUE (EBLUE) for β and the empirical BLUP (EBLUP) for u by plugging the estimated components into the formulas of the BLUEs, or BLUPs (Hobza et al. 2021: p. 189). The empirical versions of the best linear unbiased estimator for β and of the best linear unbiased predictor for u are

$$\hat{\beta} = (X^{'} \hat{V}^{-1} X)^{-1} X^{'} \hat{V}^{-1} y \quad \text{and} \quad \hat{u} = \hat{V}_u Z^{'} \hat{V}^{-1} (y - X\hat{\beta}),$$

with

$$\hat{V} = Z \hat{V}_u Z^{'} + \hat{V}_\epsilon, \qquad \hat{V}_u = \hat{\sigma}_u^2 I_D, \qquad \hat{V}_\epsilon = \hat{\sigma}_\epsilon^2 W^{-1}.$$

The variance components can be estimated using maximum likelihood (ML), restricted maximum likelihood (REML) or H3 methods (Hobza et al. 2021: p. 157 ff.). However, a more detailed description of these methods will not be given in this thesis. Instead, BLUPs and EBLUPs for area level models are considered now.

Best linear unbiased predictors under area level models

In this subsection, the results of the EBLUP estimation are applied to the basic area level model introduced as the Fay-Herriot model. While the previous subsection on BLUPs for nested error regression models at unit level mainly used results from Hobza et. al (2021: p. 155 ff.), this section primarily refers to Rao and Molina (2015: p. 123 ff.). At this point it should be repeated that the basic area level model is given by $\hat{\theta}_d = x_d^T \beta + z_d u_d + \epsilon_d$. The area effects $u_d \overset{iid}{\sim} (0, \sigma_u^2)$ are again independent of the sampling errors $\epsilon_d \overset{ind}{\sim} (0, \sigma_{\epsilon_d})$. To avoid the double index in the variance expression of the sampling errors, σ_{ϵ_d} is from now on denoted as ψ_d. Thus, the variance-covariance matrix of $\hat{\theta}$ is

$$V_d = \psi_d + \sigma_u^2 z_d^2.$$

The best linear unbiased predictor estimator for the target parameter $\mu_d = \theta_d = x_d^T \beta + z_d u_d$ is then given by (Rao, Molina 2015: p. 124):

$$\tilde{\theta}_d = x_d^T \tilde{\beta} + \gamma_d^v (\hat{\theta}_d - x_d^T \tilde{\beta}) \tag{5.9}$$

$$= \gamma_d^v \hat{\theta}_d + (1 - \gamma_d^v) x_d^T \tilde{\beta}, \tag{5.10}$$

with

$$\gamma_d^v = \frac{\sigma_u^2 z_d^2}{\psi_d + \sigma_u^2 z_d^2}.$$

Again, $\tilde{\beta}$ is the best linear unbiased estimator (BLUE) for β and is in this case composed of:

$$\tilde{\beta} = \tilde{\beta}(\sigma_u^2) = \left(\sum_{d=1}^{D} \frac{x_d x_d^T}{\psi_d + \sigma_u^2 z_d^2} \right)^{-1} \left(\sum_{d=1}^{D} \frac{x_d \hat{\theta}_d}{\psi_d + \sigma_u^2 z_d^2} \right).$$

Similar to 5.8, one also sees the composite form of the BLUP estimator in 5.10. It is again a weighted average of the direct estimator $\hat{\theta}_d$ and the regression synthetic estimator $x_d^T \tilde{\beta}$ including the BLUE $\tilde{\beta}$, where the weight γ_d^v in 5.10 refers to the ratio of the model variance $\sigma_u^2 z_d^2$ to the total variance $\psi_d + \sigma_u^2 z_d^2$. Thus, the BLUP form of the Fay-Herriot estimator also takes into account the between-area variation in relation to the precision of the direct estimator. This form implies that it adjusts the regression synthetic estimator $x_d^T \tilde{\beta}$ to take care of possible model deviations. The synthetic estimator is particularly emphasised when the model variance $\sigma_u^2 z_d^2$ is relatively small, since in this case γ_d^v takes on a small value. On the other hand, if the design variance ψ_d is small, γ_d^v becomes relatively large and the direct estimator $\hat{\theta}_d$ is weighted more. For huge samples, the sampling variance tends to zero $\psi_d \to 0$ and simultaneously $\gamma_d^v \to 1$, which is why $\tilde{\theta}_d$ is design-consistent. The design bias of the BLUP estimator is given by:

$$B_\pi(\tilde{\theta}_d) \approx (1 - \gamma_d^v)(x_d^T \beta^* - \theta_d),$$

with $\beta^* = \mathbb{E}_2(\tilde{\beta})$ as the conditional expectation of $\tilde{\beta}$ given θ. For the reason that $\gamma_d^v \to 1$, it follows that the design bias in relation to θ_d approximates zero for large samples (Rao, Molina 2015: p. 124 f.).

After the bias, now the mean squared error for the BLUP estimator is considered, which is given by:

$$MSE(\tilde{\theta}_d) = \mathbb{E}(\tilde{\theta}_d - \theta_d) = g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2). \tag{5.11}$$

The expression in 5.11 consists of two parts, that is $g_{1d}(\sigma_u^2)$ and $g_{2d}(\sigma_u^2)$. The first part is determined by

$$g_{1d}(\sigma_u^2) = \frac{\sigma_u^2 z_d^2 \psi_d}{(\psi_d + \sigma_u^2 z_d^2)} = \gamma_d^v \psi_d,$$

and for the second part applies

$$g_{2d}(\sigma_u^2) = (1-\gamma_d^v)^2 x_d^T \left(\sum_{d=1}^{D} \frac{x_d x_d^T}{(\psi_d + \sigma_u^2 z_d^2)} \right)^{-1} x_d.$$

The influence of $g_{1d}(\sigma_u^2)$ and $g_{2d}(\sigma_u^2)$ on the MSE is also weighted by γ_d^v. Focusing only on the first part $g_{1d}(\sigma_u^2) = \gamma_d^v \psi_d$, one sees that the design variance of the direct estimator $\mathbb{V}(\hat{\theta}_d) = \psi_d$ contributes to the MSE with the weight γ_d^v. If γ_d^v is small, the BLUP estimator leads to a large efficiency gain, since the variance of the model error $z_d u_d$ is small in relation to the total variance (Rao, Molina 2015: p. 125 f.).

As with the nested error regression model, the variance component σ_u^2 of the basic area level model is unknown in practice and is replaced by the estimate $\hat{\sigma}_u^2$. This makes the BLUP estimator become the empirical BLUP again, which for the basic area level model is given by

$$\hat{\theta}_d^{EBLUP} = \hat{\gamma}_d^v \hat{\theta}_d + (1-\hat{\gamma}_d^v) x_d^T \hat{\beta}, \tag{5.12}$$

where $\hat{\gamma}_d^v$ and $\hat{\beta}$ are the values of γ_d^v and β, when σ_u^2 is replaced by $\hat{\sigma}_u^2$, thus

$$\hat{\gamma}_d^v = \frac{\hat{\sigma}_u^2 z_d^2}{\psi_d + \hat{\sigma}_u^2 z_d^2} \qquad \text{and} \qquad \hat{\beta} = \tilde{\beta}(\hat{\sigma}_u^2) = \left(\sum_{d=1}^{D} \frac{x_d x_d^T}{\psi_d + \hat{\sigma}_u^2 z_d^2} \right)^{-1} \left(\sum_{d=1}^{D} \frac{x_d \hat{\theta}_d}{\psi_d + \hat{\sigma}_u^2 z_d^2} \right),$$

which are plugged into formula 5.12 (Rao, Molina 2015: p. 126). The estimation of the variance component σ_u^2 is very complex in practice. Here a method is sketched to obtain a moment estimator $\hat{\sigma}_{um}^2$. It applies that:

$$\mathbb{E}\left(\sum_{d=1}^{D} \frac{(\hat{\theta}_d - x_d^T \tilde{\beta})^2}{(\psi_d + \sigma_u^2 z_d^2)} \right) = \mathbb{E}(a(\sigma_u^2)) = m - p.$$

One obtains a solution $\hat{\sigma}_{um}^2$, if one solves iteratively

$$a(\sigma_u^2) = m - p$$

keeping $\hat{\sigma}_{um}^2 = 0$, when no positive solution $\tilde{\sigma}_{um}^2$ exists. For a possible iterative solution, one chooses a starting value $\sigma_u^{2(0)}$ and then defines

$$\sigma_u^{2(k+1)} = \sigma_u^{2(k)} + \frac{1}{a_*^{'}\left(\sigma_u^{2(k)}\right)} \left(m - p - a\left(\sigma_u^{2(k)}\right)\right),$$

with

$$a_*^{'}(\sigma_u^2) = -\sum_{d=1}^{D} z_d^2 \frac{(\hat{\theta}_d - x_d^T \tilde{\beta})^2}{(\psi_d + \sigma_u^2 z_d^2)^2}$$

as an approximation to the derivative $a^{'}(\sigma_u^2)$. This iterative method achieves rapid convergence and was also used by Fay and Herriot in estimating log per capita income for small places. Alternatives to this technique would be, for example, a simple moment estimator $\hat{\sigma}^2_{us}$ or the Fisher-scoring algorithm for ML or REML estimation of σ^2_u (Rao, Molina 2015: p. 126 ff.).

Now that the concept of (E)BLUP estimation for the basic area level model and an approach for estimating the variance component σ^2_u have been presented, two further aspects of BLUP estimation will be discussed. First, it is analysed what happens if not all areas are sampled. Since no sample information is available in the non-sampled areas, no direct estimators can be calculated, which are required in the first part of the (E)BLUP formula 5.10, respectively 5.12. In these areas, the regression synthetic estimator of θ_d based on the covariates, observed from the non-sampled areas, is used instead

$$\hat{\theta}^{RS}_l = x_l^T \hat{\beta}, \qquad\qquad \hat{\beta} = \tilde{\beta}(\hat{\sigma}^2_u),$$

where $l = m+1, \ldots, D$ are the non-sampled areas (Rao, Molina 2015: p. 126).

This section is concluded with a compromise EBLUP estimator $\hat{\theta}^{EBLUP}_{dc}$ proposed by Fay and Herriot (1979: p. 271). For this estimator the following applies with a specified constant c:

$$\hat{\theta}^{EBLUP}_{dc} = \begin{cases} \hat{\theta}^{EBLUP}_d - c\sqrt{\psi_d}, & \text{if } \hat{\theta}^{EBLUP}_d < \hat{\theta}_d - c\sqrt{\psi_d} \\ \hat{\theta}^{EBLUP}_d, & \text{if } \hat{\theta}^{EBLUP}_d \in [\hat{\theta}_d - c\sqrt{\psi_d}; \hat{\theta}_d + c\sqrt{\psi_d}] \\ \hat{\theta}^{EBLUP}_d + c\sqrt{\psi_d}, & \text{if } \hat{\theta}^{EBLUP}_d > \hat{\theta}_d - c\sqrt{\psi_d} \end{cases}$$

For the constant, the value $c = 1$ is typically chosen.

Now that the Battese-Harter-Fuller estimator and the Fay-Herriot estimator have been presented in detail and their important properties as the (empirical) best linear unbiased predictor have been discussed, this chapter concludes with a brief discussion of other model-based small area estimators.

5.6 Further model-based small area estimators

For the model-based estimation of small area values, there are many other possibilities besides the BHF and the FH estimator. Since these were introduced in such detail, the alternatives are only presented very briefly in each case. It will be started with area level temporal linear mixed models.

Area level temporal linear mixed models

Until now, the principle of borrowing strength has only been applied spatially to the areas. In the presence of temporal data, it is a good possibility to additionally borrowing strength from time. Thus, models can be used that borrow information across areas and over time, which also include previously unexplained area-time variation. Here, two temporal extensions of the Fay-Herriot model are briefly given, one with independent domain-time random effects and one with a correlation between the time points within the areas. The area level linear mixed model with independent time effects is given by

$$y_{dt} = x_{dt}\beta + u_{1,d} + u_{2,dt} + \epsilon_{dt} \tag{5.13}$$

for $d = 1, \dots D$ areas and $t = 1, \dots, T$ time periods as well as $u_{1,d} \sim N(0, \sigma_1^2)$, $u_{2,dt} \sim N(0, \sigma_2^2)$ and $\epsilon_{dt} \sim N(0, \sigma_{dt}^2)$. Also for this model, best linear unbiased estimator and predictor formulas for β and u can be given for known variance parameters σ_1^2 and σ_2^2. The equation for the area level model with correlated time effects takes the same form as 5.13. However, an autocorrelation is assumed for the domain-time random effects u_2. Thus, the variance of u_2 is given by $V_{u2} = \sigma_2^2\Omega(\phi)$ with $\Omega(\phi) = \underset{1 \leq d \leq D}{diag} (\Omega_d(\phi))$ (Hobza et al. 2021: p. 461 ff.).

Spatial area models

In addition to the temporal influence, it also makes sense in some applications to consider the spatial structures and dependencies in the modelling. In the basic BHF and FH models it is assumed that the area random effects u_d are independent of each other. In some cases, however, it is useful to include the spatial dependencies via correlations in the area random effects. For this, neighbourhoods between the areas are defined. In the geographical context, for example, all areas that have a common border with d belong to the neighbourhood of area d. All neighbours of an area d are contained in the set A_d. Based on this, a conditional autoregressive spatial model assumes that the conditional distribution of $z_d u_d$, given the area effects of the other areas $\{u_l : l \neq d\}$, is given by:

$$z_d u_d \mid \{u_l : l \neq d\} \sim N\left(\rho \sum_{l \in A_d} q_{dl} z_l u_l, z_d^2 \sigma_u^2\right).$$

In this case, the q_{dl} are known constants for which $q_{dl} z_l^2 = q_{ld} z_d^2$ as well as $q_{dl} = 0$ if the area l is not a neighbour of the area d and $\delta = (\rho, \sigma_u^2)$ is the unknown parameter vector (Rao, Molina 2015: p. 86 f.). This information for the area-specific random effects is taken into account in the small area models, in order to then also include the spatial structures in the modelling.

Two-fold subarea models

Another case of small area estimation that differs from the classic BHF and FH models is the subdivision of the areas into further subareas. Here it has again to be distinguished between the area level and the unit level. At area level, each area d is partitioned into N_d subareas. Here one is interested in both the area means θ_d and the subarea means θ_{dj} with $j = 1, \ldots, N_d$ and $d = 1, \ldots, D$. For this a linking model is given by $\theta_{dj} = x_{dj}^T\beta + u_d + v_{dj}$ with x_{dj} as subarea level auxiliary information and the area effects $u_d \sim N(0, \sigma_u^2)$ are independent of the subarea effects $v_{id} \sim N(0, \sigma_v^2)$. Then n_d of the N_d subareas are sampled in area d and $\hat{\theta}_{dj}$ is a direct estimator of the subarea mean θ_{dj} for a selected sample of n_{dj} units in the subarea dj. Thus, the sampling model in combination with the linking model gives the two-fold subarea level model:

$$\hat{\theta}_{dj} = x_{dj}^T\beta + u_d + v_{dj} + \epsilon_{dj}.$$

This model can by used to estimate the area means θ_d as well as the subarea means θ_{dj} by borrowing strength from the concerning areas and subareas (Rao, Molina 2015: p. 88).

Similar is the situation at unit level. Here the dth area is divided into M_d clusters or primary units where the ith primary unit in the dth area contains N_{di} subunits or elements. For this, let (y_{dij}, x_{dij}) denote the y and x-values of the jth element in the ith primary unit of the dth area. Within this framework, a sample s_d of m_d clusters is selected in area d. Then, from the ith sampled cluster, a subsample s_{di} of n_{di} elements is sampled. Thus, the two-fold nested error regression model at unit level is given by

$$y_{dij} = x_{dij}^T\beta + u_d + v_{di} + \epsilon_{dij}$$

with area effects $u_d \overset{iid}{\sim} (0, \sigma_u^2)$, cluster effects $v_{di} \overset{iid}{\sim} (0, \sigma_v^2)$ and residual errors ϵ_{dij} (Rao, Molina 2015: p. 89 f.). With this, the two-fold model can also be used at unit level.

Multivariate unit level and area level models

So far in this chapter, the parameter of the target variable θ_d has always been composed of only one characteristic. Now the multivariate case is considered, when a model is to be used to estimate several variables of interest at once. At unit level, it is still assumed that unit-specific auxiliary data x_{di} are available for all population elements i in each small area d. Only, in contrast to the formulae in section 5.2, y_{di} is now a $r \times 1$ vector with r variables of interest. Then the nested error regression model results in:

$$y_{di} = Bx_{id} + u_d + \epsilon_{di}.$$

In this case B is a $r \times p$ matrix of the regression coefficients and for the $r \times 1$ vectors of the area effects u_d holds $u_d \overset{iid}{\sim} (0, \Sigma_u)$. For the vectors of errors applies $\epsilon_{di} \overset{iid}{\sim} (0, \Sigma_\epsilon)$. For a large population size N_d the target parameters with the vectors of area means $\bar{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} y_{di}$ can be approximated by $\mu_d = B\bar{X}_d + u_d$. The multivariate case is able to lead to a more efficient estimation if the correlations between the components of y_{di} are taken into account (Rao, Molina 2015: p. 89 f.).

The multivariate case can also be considered for the Fay-Herriot model at area level. Here, a $r \times 1$ vector with the area characteristics of interest $\theta_d = (\theta_{d1}, \dots, \theta_{dr})^T$ is regarded with $\theta_{di} = g_i(\bar{Y}_{di})$ and $\bar{Y}_{di}$ as the dth small area mean for the ith variable. In the multivariate sampling model $\hat{\theta}_d = \theta_d + \epsilon_d$, the sampling errors are independent r-variate normal $\epsilon_d \sim N_r(0, \Psi_d)$. Moreover, θ_d is linked to the area-specific auxiliary data $\{x_{di}\}$ via the linear model $\theta_d = X_d\beta + u_d$ with the area-specific random effects $u_i \sim N_r(0, \Sigma_u)$, so that the multivariate mixed linear model at area level is given by:

$$\hat{\theta}_d = X_d\beta + u_d + \epsilon_d.$$

Here β is an rp vector of the regression coefficients and X_d is an $r \times rp$ matrix of the auxiliary data. This model can also lead to more efficient estimators of the small area means when correlations between the components of $\hat{\theta}_d$ are taken into account (Rao, Molina 2015: p. 81 f.).

Generalized linear mixed area models

The previous models in this chapter, especially the nested error regression model for the BHF estimator and the Fay-Herriot model, are linear models. However, generalised linear models can also be used for small area estimation. For example, to estimate the proportion of a binary variable of interest $\bar{Y}_d = P_d = \sum_{i=1}^{N_d} \frac{y_{di}}{N_d}$, a logistic regression model with random area-specific effects can be used:

$$\hat{p}_{di} = \text{logit}(p_{di}) = \log\left(\frac{p_{di}}{1 - p_{di}}\right) = x_{di}^T\beta + u_d.$$

A model based estimator for the area proportion is the given by:

$$\hat{P}_d = \frac{1}{N_d}\left(\sum_{i \in s_d} y_{di} + \sum_{i \in r_d} \hat{p}_{di}\right).$$

A similar approach was used in the previous chapter for the mixed logistic generalised regression estimator. If the target variable is a categorical variable with K categories, a multinomial logistic model with random area-specific effects is used instead of the logistic model. In other cases, further distributions can equally be used, such as the Poisson distribution for the occurrence of rare events or the gamma distribution as well as the exponential distribution. Splines can also be used for even more flexible modelling (Rao, Molina 2015: p. 92 ff.).

This chapter, in addition to the detailed presentation of the Battese-Harter-Fuller and the Fay-Herriot estimator, has shown how diverse the model-based modelling of small area estimation can be. In addition to the possibilities presented here, there are of course many other approaches that cannot even be discussed here. Instead, georeferencing and small area estimation will be analysed in the light of deficient data in the next chapter.

Chapter 6

Georeferencing and small area estimation in the light of deficient data

Meanwhile, different types of deficit data as well as georeferencing and small area estimation including their central estimators have been introduced in this thesis. This chapter focuses on the interaction of these components and examines how deficit data affect georeferencing and small area estimation. It is started with georeferencing and the influence of deficient data on it.

6.1 Georeferencing with deficient data

It has already been pointed out in some parts of this thesis that small area estimation is very often based on georeferenced data. This means that deficient georeferenced data also have an impact on small area estimation. In this section some cases are considered how georeferencing can lead to deficient data. One danger with georeferencing is informal georeferences. Instead of exact coordinates, for example, respondents' addresses are often requested via street names in oral interviews. Thus it can happen that even if the address is recorded correctly, especially in large cities, several streets have the same name or a street has several spellings (Hackeloeer et al. 2014: p. 65). This does not result in missing data, but it can generate incorrect data, which leads to biases in the analysis. In the case of two identical street names within a city, it is very likely that they are in different quarters and using quarters as small areas will lead to biased analyses. Another problem in this regard are cities where several streets have a similar name. This can quickly lead to confusion, especially during oral surveys. Therefore, whenever possible, street names and other semantic properties should never be used as the single source of identification of spatial information (Hackeloeer et al. 2014: p. 65).

But non-semantic georeferencing can also lead to deficient data. Thus, a spatial dataset whose georeferences were generated via coordinates can also be afflicted with errors, especially measurement errors. For this, it is useful to look at the terms precision, bias and accuracy in this context. Precision relates to the dispersion of random position errors and is typically indicated by a standard deviation or variance. Bias, on the other hand, is related to systematic errors and is generally expressed by an average error, which should optimally be zero. Accuracy is related to both precision and bias and indicates how close the features on the map are to their actual position on the ground (Ribeiro et al. 2014: p. 3). All these terms play a role in the process introduced in the third chapter as geocoding. This is the transformation step needed to convert georeferenced data into a desired reference system. Errors can occur in the assignment of addresses, so-called address georeferencing, to a specific position on earth, which can be larger or smaller depending on the method used. One source of error is positional errors, which occur when maps are created and depend on the scale used, i.e. the relationship between the distance on the map and the real distance. The use of maps with errors can lead to a reduced accuracy of the georeferenced data. Insufficient positional accuracy could interfere with cluster detection and impact the magnitude of the regression coefficients. If some objects cannot be assigned coordinates during georeferencing, this can reduce the statistical power and generate a bias due to so-called non-random missingness (Ribeiro et al. 2014: p. 3). As was seen in the second chapter in the part on missing data, dealing with missing not at random is very complicated. The points mentioned here show that in the statistical analysis of spatial data, as is the case with small area estimation, sources of error not arising from the statistical survey process also have to be taken into account. There are various possibilities for the geocoding process. Geographical information systems (GIS), Google Earth or so-called global positioning system (GPS) receivers can be used, although all methods have sources of error. With GIS, addresses are placed in a corresponding street segment and then assigned coordinates. Here it can happen that the assignment does not work and that errors occur as a result. With GPS receivers, positional errors of ten to 20 metres occur very frequently (Ribeiro et al. 2014: p. 6 ff.). For large spatial areas, this is less of a problem in the analysis, but for smaller spatial areas, such as grids of 100 by 100 metres, positional errors of over ten metres can lead to major problems. Especially for individuals at the edge of areas, this can lead to them being incorrectly assigned and a grid-based evaluation, for example, then leads to biased results. In the analysis by Ribeiro et al. (2014: p. 10), in which addresses were to be assigned to the various census tracts, the misclassification rate ranged from 14 to 38 percent, depending on the georeferencing method used (GIS, GPS, Google Earth). This makes it clear how diverse the sources of error are in spatial analyses and through which types deficient data can arise.

The next step is again to examine rather statistical sources of error and deficient data in the grid-based evaluation, whose diverse areas of application have already been presented in the third chapter on georeferencing. Likewise, advantages and disadvantages of grid-based evaluation have already been presented. At this point, disadvantages are analysed again in relation to deficient data and results from other chapters are included. In the second chapter of this thesis, confidentiality methods were already considered in the context of deficient data. Especially in official statistics, the protection of data is a high good and must not be neglected under any circumstances. This point must also be taken into account in grid-based georeferencing. Especially in grids with a small width, such as 100 by 100 metre grids, it can happen that in some grids there are no observations at all or only very few. In this case, the data are particularly worthy of protection and it may happen that no data are published for these grids and they are therefore missing. However, other data confidentiality methods from the second chapter can also be used, which all lead to deficient data. This might not have happened if other spatial layers, such as administrative areas, had been used, as these are often much larger than grids and therefore usually contain more observations. In return, the small-scale analysis potential for such areas is not as high as for the small grids. This is a trade-off that one is often confronted with in such cases and which is assessed differently from case to case. The Census Atlas (Neutze 2015: p. 65) also has to deal with similar problems. Here there is a high number of grid cells that are not allowed to contain any data for reasons of non-disclosure and are therefore not available for analysis.

In addition to the missing or altered data, small grid cells with few observations can lead to deficient results in another respect, namely in the estimation. Thus, in the previous two chapters, a large set of estimators were considered, some of which show big problems of precision with small sample sizes. One example is the Horvitz-Thompson estimator, which has an enormously large variance in such cases. This will also become clear in the next chapter of the practical analysis in the small Munich city district quarters. Although some estimators were presented that also provide estimates in unsampled areas with the help of auxiliary variables, there is often no auxiliary data available at this level, especially in the approach with grids of a small size that has not been used for long. In this case, the number of estimators is severely limited and many of those available, especially the direct estimators, thus lead to problems. These are two problematic aspects that should not be ignored in addition to the many advantages of grid-based evaluation presented in the third chapter. In addition to the relation of deficit data and georeferencing, the remainder of this chapter considers small area estimation in the context of deficit data.

6.2 Small Area Estimation in the light of deficient data

It should be mentioned at the outset that due to the many types of deficient data, there are a variety of ways in which deficient data affects small area estimation. This section presents some of these and also discusses solutions on how small area estimation deals with them. The first step is to start with outliers and robust small area estimation.

6.2.1 Robust small area estimation

In the chapter on deficient data, outliers were classified as particularly influential observations. These then have a particularly high weight when considering the arithmetic mean, for example. Outliers also have a problematic effect on the assumption of normal distributions, as in many of the estimators presented so far. Therefore, robust small area estimation is presented as a way to counteract these problems. A proposal for dealing with outliers in direct small area estimators is presented by Chambers (1986: p. 1064 ff.). For this, the population is divided into the sampled part s and the non-sampled part r. A regression estimator $\hat{Y}_d = \sum_{i \in s_d} y_{di} + \sum_{i \in r_d} (\hat{\beta} x_{di})$ is considered with $y = \beta x_{di} + \epsilon_{di}$. A more robust version of this is given when ϵ_{di} assumes the mixture model $\epsilon_{di} = (1 - \delta_i)\epsilon_{1di} + \delta_i \epsilon_{2di}$ with δ_i as indicator with $\pi_i = P(\delta_i = 1)$. Moreover, the variance of ϵ_{2di} is much larger than that of ϵ_{1di}. Thus, the mixture model takes into account that only a few y values are outliers.

For the mixed models used by the Battese-Harter-Fuller estimator and the Fay-Herriot estimator, among others, Sinha and Rao (2009: p. 386 ff.) present a robust approach. Here, the area model is fitted with $y_d \sim N(X_d\beta, V_d)$ where $V_d = V_d(\theta)$ depends on the variance components θ. The maximum likelihood estimators for β and θ are considered, which are obtained by solving the following ML equations.

$$\sum_{d=1}^{D} X_d^T V_d^{-1}(y_d - X_d\beta) = 0 \tag{6.1}$$

$$\sum_{d=1}^{D} \left\{ (y_d - X_d\beta)^T V_d^{-1} \frac{\partial V_d}{\partial \theta_l} V_d^{-1} - tr\left(V_d^{-1} \frac{\partial V_d}{\partial \theta_l} \right) \right\} = 0,\ l = 1, \dots, q \tag{6.2}$$

An indicator for outliers are large differences for some components of the fitted vector $\hat{y}_d = X_d\hat{\beta}$ from the corresponding observed values y_i. To better deal with the outliers, more robust versions of the ML equations in 6.1 and 6.2 are used instead:

$$\sum_{d=1}^{D} X_d^T V_d^{-1} U_d^{1/2} \psi(r_d) = 0 \tag{6.3}$$

$$\phi_l(\theta) = \sum_{d=1}^{D} \left\{ \psi^T(r_d) U_d^{1/2} V_d^{-1} \frac{\partial V_d}{\partial \theta_l} V_d^{-1} U_d^{1/2} \psi(r_d) - tr\left(K_d V_d^{-1} \frac{\partial V_d}{\partial \theta_l} \right) \right\} = 0 \tag{6.4}$$

Here, $r_d = U_d^{1/2}(y_d - X_d\beta)$ with U_d as a diagonal matrix with diagonal elements U_{di} which are equal to the diagonal elements of the covariance matrix V_d. The matrix K_d is also a diagonal matrix for which holds $K_d = cI_d$ with $c = \mathbb{E}(\psi_b^2(r))$ where r follows a standard normal distribution and the matrix I_d represents the identity matrix of the same order as V_d (Sinha, Rao 2009: p. 386).

To obtain the robust estimators for β and θ, the equations 6.3 and 6.4 are solved iteratively. One way is to use an adopted Newton-Raphson algorithm. With a first-order Taylor series expansion around β_0 an approximation of the left side of 6.3 is given as:

$$\sum_{d=1}^{D} X_d^T V_d^{-1} U_d^{1/2} \psi(r_d(\beta)) \approx \sum_{d=1}^{D} X_d^T V_d^{-1} U_d^{1/2} \psi(r_d(\beta_0)) - \sum_{d=1}^{D} X_d^T V_d^{-1} D_d(\beta_0) X_d(\beta - \beta_0),$$

with $D_d(\beta)$ as a diagonal matrix where the *ith* diagonal element is $D_{di} = \psi_b^{'}(r_{di}) = (\partial/\partial r_{di})\psi_b(r_{di})$. The Huber ψ-function is defined as $\psi_b^{'}(r_{di}) = 1$ if $|r_{di}| \leq b$ and $\psi_b^{'}(r_{di}) = 0$ otherwise. Based on this, an iterative equation for β is given as:

$$\beta^{(m+1)} = \beta^{(m)} + \left\{ \sum_{d=1}^{D} X_d^T V_d^{-1} D_d(\beta^{(m)}) X_d \right\}^{-1} \left\{ \sum_{d=1}^{D} X_d^T V_d^{-1} U_d^{1/2} \psi(r_d(\beta^{(m)})) \right\}. \tag{6.5}$$

An iterative equation for θ is obtained similarly to this approach by solving the robust ML equation in 6.4 using the Newton-Raphson algorithm:

$$\theta^{(m+1)} = \theta^{(m)} - (\phi^{'}(\theta^{(m)}))^{-1} \phi(\theta^{(m)}) \tag{6.6}$$

The overall algorithm for robust estimation of β and θ can be described as follows. First, initial values $\beta^{(0)}$ and $\theta^{(0)}$ are chosen and m is set to $m = 0$. Using the equations 6.5 and 6.6, the values for $\beta^{(m+1)}$ and $\theta^{(m+1)}$ are calculated and then m is set to $m = m + 1$. This is done until the algorithm converges and the estimators are the RML estimators $\hat{\beta}_M$ and $\hat{\theta}_M$ of β and θ respectively. These robust estimators $\hat{\beta}_M$ and $\hat{\theta}_M$ are then used to calculate the area-specific random effects $\hat{u}_{dM}$ for the individual areas. These results can then be taken to generate the robust empirical best linear unbiased predictions (REBLUP) for the total values Y_d of the individual areas. A special bootstrap procedure is used to calculate the mean square prediction error for the robust estimators. (Sinha, Rao 2009: p. 386 ff.).

The proposal by Sinha and Rao just presented in detail is a possibility for robust small area estimation when outliers occur. There are also other approaches. As already mentioned, outliers have a particular impact on the arithmetic mean. The median is less affected. This only looks at the value for which 50 percent of the data are smaller and 50 percent of the data are larger. It makes no difference to the median how much the observations at the edges of the distribution deviate from the rest. Furthermore, the interquartile range would be an alternative for a robust variability measure. This shows that even small changes can have a positive effect on the robustness of the small area estimation. In the practical part of this thesis, the use of the median is later compared with the use of the arithmetic mean. At this point, the influence of missing data on small area estimation and ways of dealing with it are now examined.

6.2.2 Small area estimation with missing data

Two approaches for dealing with missing data are considered in this section. On the one hand, an approach is presented with imputation procedures, by which the missing data are eliminated and the estimation takes place on complete data, and on the other hand, a procedure is looked at in which the missing data are taken into account in the modelling.

Imputation methods

The first way presented for handling missing data as a form of deficient data are imputation methods. Basically, one can distinguish between single imputation and multiple imputation. With single imputation, exactly one value is substituted for each missing value. One possibility for this is mean imputation. Here, all missing values of a variable are replaced by the arithmetic mean of the observed values of this characteristic. Thus, the mean value of the distribution of this variable remains unchanged. However, by inserting the arithmetic mean, many values are placed in the centre of the distribution and the variance of the characteristic is clearly underestimated. One way to counteract this aspect a little is to group them into classes. For this, the observations of the data set are divided into groups according to other variables and the group mean is then used for all missing values (Little, Rubin 2002: p. 61 f.). A further possibility for single imputation is regression imputation. For each variable with missing data, a model is set up with the other characteristics serving as covariates. The missing values are then fitted as a prediction using the estimated parameters. In order not to systematically underestimate the variance in the data, a stochastic component can also be added. The imputed values then represent random drawing from a predictive distribution of plausible values and no longer all originate from the centre of the distribution as, for example, with mean imputation (Little, Rubin 2002: p. 64 ff.). For the elimination of missing data, the so-called hot deck imputation can also be used. The missing values of a unit are replaced by the observations of a similar unit. Based on the observed variables, distance matrices are created between the units and the unit with the smallest distance is selected to replace the missing value (Little, Rubin 2002: p. 60).

In all these presented methods of single imputation, the imputed value is considered to be known and fixed. This means that the variance of the modelling of the missing values is lost. To counteract this aspect, the methods of single imputation should be used in the context of multiple imputation. For this, the data set with the missing data is copied so that it exists at least $D \geq 2$ times. One of the single imputation methods is then applied to each of these D data sets. Then the desired analyses are carried out on each of the complete D data sets. In the context of small area estimation, for example, the Horvitz-Thompson estimators, the Battese-Harter-Fuller estimators or the Fay-Herriot estimators are calculated on each data set. The D results are then averaged. For this purpose, the mean value of the D runs is used for the parameter of interest, that is, in the context of small area estimation, the value of the estimator under consideration. The variance of the parameter, which arises in the context of the multiple imputation process, is composed of the average of the within-imputation variance and a between-imputation component (Little, Rubin 2002: p. 85 ff.). Imputation is only one way of dealing with missing values in the context of small area estimation. In any case, it must be noted that there is no certainty as to how far the imputed values differ from the real true value. In addition, one should always consider which missing data mechanism is present before applying imputation mechanisms, although this can never be precisely determined in practice. Another approach to dealing with missing values is presented below.

Cautious modelling of missing data in small area estimation

The second possibility presented here has been proposed by Plass, Omar and Augustin (Plass et al. 2017: p. 253 ff.) and is called cautious modelling of missing data in small area estimation. Here it is tried not to make too strong assumptions about the missing data. It is sufficient if parameters are only partially identified and imprecise results are accepted that are credible for this, which would no longer be the case with too strong assumptions about the absence of data. If more knowledge about missingness is available at a later stage, the imprecise estimators can still be made more precise. Since no strict assumptions are made about the missingness mechanism, uncertainty due to non-response or missing data must be considered as a lack of knowledge. The cautious approach, which will be looked at in more detail below, also takes into account that nonresponse and the missing data are particularly problematic in small area estimation, since small sample sizes become even smaller here (Plass et al. 2017: p. 254).

Cautious modelling of missing data is mainly used in design-based small area estimators. In the following, an estimator is to be found that also yields credible results for the logistic generalised regression estimator

$$\hat{\pi}_d^{LGREG} = \sum_{g=1}^{G} \left(\sum_{i \in s_{d,g}} w_{di} y_{di} + \hat{\pi}^{[g]} (X_d^{[g]} - \sum_{i \in s_{d,g}} w_{di}) \right) / N_d, \qquad \hat{\pi}^{[g]} = \sum_{d=1}^{D} \sum_{i \in s_d^{[g]}} \frac{y_{di}}{n^{[g]}}$$

without making assumptions about the missingness of data. The denomination g stands for the g-th subgroup of the cross-classified categorical covariates. First, an lower bound $\underline{\hat{\pi}}_i^{SYN}$ and a upper bound $\overline{\hat{\pi}}_i^{SYN}$ are calculated for the synthetic estimator, these being for the extreme cases where either all missing values of the binary target variable take the value 0, $y_{di} = 0$, or all missing values are taken as 1, $y_{di} = 1, \forall i \in s_{d,mis}, d = 1, \ldots, D$ (Plass et al. 2017: p. 258):

$$\underline{\hat{\pi}}_i^{SYN} = \frac{1}{N} \sum_{d=1}^{D} \sum_{i \in s_{d,obs}} w_{di} y_{di}, \qquad \overline{\hat{\pi}}_i^{SYN} = \frac{1}{N} \sum_{d=1}^{D} \left(\sum_{i \in s_{d,obs}} w_{di} y_{di} + \sum_{i \in s_{d,mis}} w_{di} \right).$$

The index d for the areas is divided into d^* for the area of interest and $d \neq d^*$ for all other areas. Then a cautious form for the $\hat{\pi}_{d^*}^{LGREG}$ in the context of missing data is given by:

$$\sum_{g=1}^{G} \left(\left(\sum_{\substack{d=1 \\ d \neq d^*}}^{D} \sum_{i \in s_d^{[g]}} \frac{y_{di}}{n^{[g]}} \right) \left(X_{d^*}^{[g]} - n_{d^*}^{[g]} w_{d^*} \right) + \sum_{i \in s_{d^*}^{[g]}} \frac{y_{d^*i}}{n^{[g]}} \left(X_{d^*}^{[g]} - w_{d^*} (n_{d^*}^{[g]} + n^{[g]}) \right) \right) / N_{d^*}, \quad (6.7)$$

$$\text{with} \sum_{i \in s_d^{[g]}} \frac{y_{di}}{n^{[g]}} = \sum_{i \in s_{d,obs}^{[g]}} \frac{y_{di}}{n^{[g]}} + \sum_{i \in s_{d,mis}^{[g]}} \frac{y_{di}}{n^{[g]}} \text{ and } \sum_{i \in s_{d^*}^{[g]}} \frac{y_{d^*i}}{n^{[g]}} = \sum_{i \in s_{d^*,obs}^{[g]}} \frac{y_{d^*i}}{n^{[g]}} + \sum_{i \in s_{d^*,mis}^{[g]}} \frac{y_{d^*i}}{n^{[g]}}.$$

Now the problem is to find the values of y_{di} for the nonrespondents respectively for the missing data so that the equation 6.7 is minimized respectively maximized. Therefore an separate optimization for each subgroup $g = 1, \ldots, G$ is needed. One possibility here would again be to assume the extreme cases, that the missing values are all either equal to 0 or equal to 1 (Plass et al. 2017: p. 258). This estimator just presented is used when one does not want to make any assumptions about the absence of data at all. If, on the other hand, partial knowledge is available, partial missingness assumptions can also be included. In a practical example, it is shown that even very weak assumptions about the missingness of the data provide a much more accurate interval for the synthetic estimator and the logistic generalised regression estimator, and the lower and upper bounds of the estimators are much closer to each other than in the case where no assumptions are made about the missingness of the data (Plass et al. 2017: p. 258 ff.).

However, the advantages of this cautious approach are hardly applicable to model-based estimators such as the Fay-Herriot estimator or the Battese-Harter-Fuller estimator, since the use of random effects u_d results in analytically unsolvable integrals and the cautious likelihood approach reaches its limits (Plass et al. 2017: p. 261 f.). Therefore, in the practical analysis of this thesis, the effects of missingness on the estimators and a simple imputation method are considered in the next chapter when dealing with missing data. At this point, it will be continued with small area estimation in the context of measurement errors.

6.2.3 Small area estimation for data with measurement errors

Using auxiliary information from external datasets is considered one of the core elements of small area estimation. Often, register data is used for this purpose, which, however, does not always provide suitable auxiliary variables. Thus, alternative surveys are often used in which the auxiliary variables have measurement errors. Therefore, this section considers a version of the Fay-Herriot model that takes these measurement errors into account in the auxiliary data (Burgard et al. 2021: p. 79 ff.). This modelling approach consists of three steps and is based on the fact that the vector of true domain means of the auxiliary data is different from the associated vectors of direct estimates in a multivariate normally distributed random error. The main goal of the approach is to calculate empirical best predictors for the means of the small areas and to estimate the associated mean squared errors (Burgard et al. 2021: p. 81). The approach presents a measurement error bivariate Fay-Herriot model, where the bivariate model can be seen as a part of the multivariate area level models introduced in the last section of the previous chapter on further model-based small area estimators.

Here $\mu_d = (\mu_{d1}, \mu_{d2})$ is the bivariate vector with the variables of interest in area d and $y_d = (y_{d1}, y_{d2})$ denotes the vector of direct estimators for μ_d based on the target survey data. In the first step, the following model is set up for the unbiased direct estimators:

$$y_d = \mu_d + \epsilon_d \tag{6.8}$$

with $\epsilon_d = (\epsilon_{d1}, \epsilon_{d2}) \sim N_2(0, V_{\epsilon d})$ independent and known covariance matrices $V_{\epsilon d}$ (Burgard et al. 2021: p. 82).

The second step assumes a linear relation of the true domain characteristic μ_{dk} to $p_k + q_k$ explanatory variables, with $k = 1, 2$. For this purpose, $\tilde{x}_{dk} = (\tilde{x}_{dk1}, \ldots, \tilde{x}_{dkp_k})$ is a vector with the true aggregated values of p_k explanatory variables for μ_{dk} and $\tilde{X}_d = \text{diag}(\tilde{x}_{d1}, \tilde{x}_{d2})$ is a $2 \times p$ block-diagonal matrix with $p = p_1 + p_2$. The vector $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{kp_k})$ contains the regression parameters for μ_{dk} with $\lambda = (\lambda_1, \lambda_2)$. $z_{dk} = (z_{dk1}, \ldots, z_{dkq_k})$ is a vector with the true aggregated values of q_k explanatory variables for μ_{dk} and $Z_d = \text{diag}(z_{d1}, z_{d2})$ is a $2 \times q$ block-diagonal matrix with $q = q_1 + q_2$. The vector $\eta_k = (\eta_{k1}, \ldots, \eta_{kq_k})$ contains the regression parameters for μ_{dk} with $\eta = (\eta_1, \eta_2)$. Based on this a linking model for the true area characteristic is given by

$$\mu_d = Z_d\eta + \tilde{X}_d\lambda + u_d, \quad u_d = (u_{d1}, u_{d2}) \sim N_2(0, V_{ud}) \tag{6.9}$$

with the vectors u_d independent of ϵ_d. For the purpose of this approach, it is assumed that the true values of $\tilde{x}_{dk}$ are unknown and predicted from independent data sources like administrative registers or other surveys where the sample sizes are larger than the target survey. The random measurement error vectors are defined as $v_{dk} = (v_{dk1}, \ldots, v_{dkp_k})$ where the vectors $v_d = (v_{d1}, v_{d2})$ are independent with $v_d \sim N_p(0, \Sigma_d)$ (Burgard et al. 2021: p. 82 f.).

In the third step, the functional measurement error model is defined as

$$\tilde{x}_{dk} = x_{dk} + v_{dk} \tag{6.10}$$

where x_{dk} is a vector with the unbiased predictors of the components of $\tilde{x}_{dk}$ and the vectors v_d and $x_d = (x_{d1}, x_{d2})$ are independent. Further two $2 \times p$ block diagonal matrices are defined as $B_d = \text{diag}(\lambda_1, \lambda_2)$ and $X_d = \text{diag}(x_{d1}, x_{d2})$ (Burgard et al. 2021: p. 83).

Combining these three steps, the measurement error bivariate Fay-Herriot model is obtained:

$$y_d = Z_d\eta + X_d\lambda + B_d v_d + u_d + \epsilon_d. \tag{6.11}$$

For the model 6.11 it is assumed that x_d, v_d, u_d and ϵ_d are independent. For the case that there are no measurement errors for the auxiliary data, the v_d's are zero and the bivariate Fay-Herriot model can be seen as a special case of 6.11. As B_d depends on λ the measurement error bivariate FH model is no longer a linear mixed model because it can not be expressed in the standard form $Y = X\beta + Zu + \epsilon$ (Burgard et al. 2021: p. 83 f.).

It is also possible to give the best predictors for the random effects v_d and u_d and the target parameter μ_d of model 6.11. In the measurement error bivariate Fay-Herriot model the best predictor for v_d is given as

$$\hat{v}_d^{BP} = \mathbb{E}(v_d \mid x_d, y_d) = \Psi_d B_d^{'}(V_{ud} + V_{\epsilon d})^{-1}(y_d - Z_d\eta - X_d\lambda)$$

with

$$\Psi_d = \left(B_d^{'}(V_{ud} + V_{\epsilon d})^{-1}B_d + \Sigma_d^{-1}\right)^{-1} = \Sigma_d - \Sigma_d B_d^{'}(V_{\lambda d} + V_{ud} + V_{\epsilon d})^{-1}B_d\Sigma_d.$$

And the best predictor of u_d in the measurement error bivariate Fay-Herriot model is

$$\hat{u}_d^{BP} = \mathbb{E}(u_d \mid x_d, y_d) = \Phi_d(V_{\lambda d} + V_{\epsilon d})^{-1}(y_d - Z_d\eta - X_d\lambda)$$

where $\Phi_d = \left((V_{\lambda d} + V_{\epsilon d})^{-1} + V_{ud}^{-1}\right)^{-1}$. Conditional on x_d, both best predictors $\hat{v}_d^{BP}$ and $\hat{u}_d^{BP}$ are unbiased. Based on this, the measurement error bivariate Fay-Herriot model best predictor for the characteristics of interest μ_d is derived as:

$$\begin{aligned}\hat{\mu}_d^{BP} &= Z_d\eta + X_d\lambda + V_{\lambda d}(V_{ud} + V_{\epsilon d})^{-1}(y_d - Z_d\eta - X_d\lambda)\\ &- V_{\lambda d}(V_{\lambda d} + V_{ud} + V_{\epsilon d})^{-1}V_{\lambda d}(V_{ud} + V_{\epsilon d})^{-1}(y_d - Z_d\eta - X_d\lambda)\\ &+ \Phi_d(V_{\lambda d} + V_{\epsilon d})^{-1}(y_d - Z_d\eta - X_d\lambda).\end{aligned}$$

Burgard et al. also present the variance of $\hat{v}_d^{BP}$ and $\hat{u}_d^{BP}$ as well as the mean squared error of $\hat{\mu}_d^{BP}$. However, this will not be discussed here. All these best predictors cannot be calculated in practice. Instead, the empirical best predictors are determined by substituting the estimated variance components:

$$\hat{\mu}_d^{EBP} = \mathbb{E}(\mu_d \mid x_d, y_d) = Z_d\hat{\eta} + X_d\hat{\lambda} + \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2)\hat{v}_d^{EBP} + \hat{u}_d^{EBP}$$

(Burgard et al. 2021: p. 84 ff.).

The presented approach has a high efficiency if the MSE of the measurement error bivariate Fay-Herriot best predictor is low in relation to the MSE of the bivariate Fay-Herriot BLUP when measurement errors occur. It is shown that the greater the measurement error variance of the auxiliary variables, the greater the efficiency gain when using the best predictor of the measurement error bivariate Fay-Herriot model. The estimation of the model parameters of the measurement error bivariate Fay-Herriot model can be done via pseudo-residual maximum likelihood or maximum likelihood (Burgard et al. 2021: p. 88 ff.). In a practical example on poverty in Spain it can be shown that by using the measurement error Fay-Herriot model, especially for small sample sizes, considerable improvements of the MSE occur in comparison to the direct estimators. This approach by Burgard et al., which is presented in detail, is only one possibility for dealing with data containing measurement errors in the context of small area estimation. Many other methods also use Bayesian procedures for this purpose. One Bayesian approach is presented below in this chapter in the following section on small area estimation for data with confidentiality methods.

6.2.4 Small area estimation for data with confidentiality methods

The final part of this chapter focuses on small area estimation in the context of data where confidentiality methods have been applied to minimise disclosure risk. It should be mentioned at the outset that this is a very diverse topic due to the many different confidentiality methods, of which a part was presented in the second chapter on deficient data. Depending on which of these methods has been used for data protection, the handling of the resulting data must also be adapted. For example, the use of suppression methods creates missing values. For this case, methods are recommended which are considered in this chapter in the part on small area estimation with missing data. If, on the other hand, noise addition is used to protect covariates with risky cells, for instance, a measurement error is added to the data and the measurement error Fay-Herriot model introduced earlier can provide a solution. As noise addition and also data swapping can be seen as methods for disclosure limitation that perturb the data, an Bayesian approach from Polettini and Arima (2015: p. 57 ff.) is looked at where measurement errors due to perturbative confidentiality methods are modelled without to strong assumption about the errors.

The following approach deals with measurement error models where the covariates are perturbed due to disclosure limitation. In this context b_d defines an area-specific covariate measured without error. On the other hand, X_d^* is used for a continuous area-specific covariate with error from an external source or as a result from perturbation of unit level data X_{di}. Z_{di}^* denotes the score of a discrete auxiliary variable with K categories after the disclosure limitation process of the original characteristic Z_{di} for unit i in area d. The vector T_d contains the frequencies of the original variable Z in area d and T_d^* denotes the frequencies of Z^* after the perturbation. The conditional distribution of T_d^* given the vector with the original values z can be approximated by a K-variate normal distribution with mean $P^{'}T_d$ where P is the perturbation matrix. Thus, the covariates $(T_{d,1}^*, \ldots, T_{d,K-1}^*)$, $T_{d,l}$ and $T_{d,l}^*$ denoting the l-th element of T_d and T_d^*, given the original scores z are modelled by a multivariate normal distribution where the mean μ_d^Z is equal to the first $K-1$ elements of $P^{'}T_d$ and the covariance matrix is $\Sigma_d^Z = \sum_{l=1}^{K-1} T_{d,l} V_{d,l}^-$, with $V_{d,l}^-$ as a $K-1 \times K-1$ submatrix where the K-th row and column of $V_{d,l}$ is dropped out (Polettini, Arima 2015: p. 62 f.).

Based on this knowledge a four-stage model is considered to deal with the perturbed data. The first stage is the model's likelihood which is given by:

$$\text{Stage 1} \;\; Y_d \mid X_d^*, Z_d^*, \theta_d, \beta, \delta \overset{ind}{\sim} N(\theta_d, \psi_d).$$

In the second stage, which is subdivided in two substages, the measurement error models for the continuous and discrete covariates are defined:

$$\begin{aligned} &\text{Stage 2.1} \;\; X_d^* \mid \theta_d, X_d \overset{ind}{\sim} N(X_d, C_d), \\ &\text{Stage 2.2} \;\; (\mathrm{T}_{d,1}^*, \dots, T_{d,K-1}^*) \overset{ind}{\sim} N_{(K-1)}(\mu_d^Z, \Sigma_d^Z). \end{aligned}$$

The third stage continues with a model for the parameter of interest θ_d as a function of covariates:

$$\text{Stage 3} \;\; \theta_d \mid X_d, T_d, \beta, \delta \overset{ind}{\sim} N(\beta_1 T_{d,1} + \cdots + \beta_{K-1} T_{d,K-1} + \beta_K X_d + \delta b_d, \;\sigma_u^2).$$

The last step is to define a uniform improper prior in order to get a proper prior density:

$$\text{Stage 4 Prior distribution } \pi(X_1, \dots, X_D, T_1, \dots, T_D, \beta, \delta, \sigma_u^2) \propto 1.$$

No strict assumptions about data perturbation are made throughout the model (Polettini, Arima 2015: p. 63 f.). Therefore, this approach is well suited for small area estimation in the context of perturbed covariates due to disclosure limitation procedures.

Now that procedures for dealing with small area estimation in the light of confidentiality methods have been introduced, the theoretical part of this work is concluded. In the next chapter, many aspects of this thesis will be practically examined using Munich population data. Aspects of this chapter will also be addressed and small area estimation will be examined in terms of some of these types of deficient data.

Chapter 7

Data analysis of the Munich municipal population data

In this chapter, the previous concepts and estimation methods from this thesis are now examined in practice with the help of Munich population data. The different estimators are compared and the influence of deficient data on small area methods is also analysed. But first, the data set used with the population data from the statistical office in Munich is presented in more detail and descriptive results are shown.

7.1 Presentation and descriptive analysis of Munich population data

The available data set of the official statistics of Munich comprises 15 characteristics for 1,578,051 residents of Munich in April 2022. Thus, almost every Munich resident who had his or her main residence in Munich in April 2022 is included in this data set. Persons without a main residence were excluded. Only a few people with main residence are missing from the dataset because they were removed using confidentiality methods before the data was passed on. These measures are explained briefly below.

7.1.1 Confidentiality methods for the population data set

As this is a microdata set, and as the second chapter highlighted how difficult these are to protect, the process of data protection on the part of the statistical office prior to data transfer is presented here in more detail. The confidentiality procedures are based on the spatial division of the city. Munich is divided into 25 city districts, which in turn are subdivided into a total of 108 city district parts. These districts can be split one level deeper into 468 city district quarters. Due to data protection concerns, all city district quarters in which fewer than 30 main residents were registered at the time were removed first. In the next step, the addresses of the persons within a quarter were permuted. This means that the address data were exchanged between the main residents of a quarter, taking care not to change the number of persons at each address. At addresses with many or at least two persons, it has happened that addresses have been exchanged between two persons at the same registration address. If this happened in a quarter with more than 40 per cent of the residents, this quarter was removed from the data set. As a result, the data now contains only 448 city district quarters instead of the original 468. With almost 1.6 million observations on 15 variables, it is nevertheless an enormously large data set, which is described in more detail below.

7.1.2 Description of the data set

In addition to administrative characteristics, the various features are mainly socio-demographic variables. As this is a German data set, the German variable name is added in brackets to the description of the individual characteristics. A precise presentation of the data set with all attributes of the individual variables can be found in table 7.1. The *reporting month* (Berichtsmonat) is the same for all persons and indicates the status of the data with April 2022. The spatial variables *city district* (Stadtbezirk), *city district part* (Stadtbezirksteil) and *city district quarter* (Stadtbezirksviertel) help with the geographical allocation of the individual units. In addition, the data set contains the *street* (Strasse), *house number* (Hausnummer) and, if available, an *alpha suffix* (Alphazusatz) for each resident of Munich. However, it must be noted here that, as described above, the addresses within the city district quarters have been permuted. Furthermore, there are the socio-demographic characteristics. Firstly, these include *age* (Lebensalter), *marital status* (Familienstand), *gender* (Geschlecht), *religious affiliation* (Religionszugehörigkeit) and *migration background* (Migrationshintergrund). The other variables give an overview of the residential history of the persons. The characteristic *duration of residence in Munich* (Wohndauer München) indicates how many years a person has had their main residence in Munich until April 2022. Two additional variables also provide information on the *immigration area* (Zuzugsbebiet). The first of these indicates the area in which the person previously lived, whereby the foreign countries are subdivided into EU and non-EU countries. Further attributes here are only resident in Munich since birth and Germany. An additional characteristic, the *immigration area in Germany* (Zuzugsgebiet Deutschland), splits the immigration areas in Germany even more finely, but has no further division for the foreign countries. In addition to these 15 variables, two further characteristics were formed. Firstly, the variable *immigration* (Zuzug), which combines the two characteristics described last and divides up the immigration areas within Germany as well as the foreign countries. It should also be mentioned that the attribute “Region 14“ for the variables immigration area in Germany and immigration in table 7.1 includes the districts of Landsberg am Lech, Fürstenfeldbruck, Starnberg, Dachau, Munich, Freising, Ebersberg and Erding. In addition, the new variable *age at immigration* (Alter beim Zuzug) was calculated from age and duration of residence in Munich and indicates how old a person was when they moved to Munich.

Variable	Description/attributes
Reporting month	April 2022
City district	Indicates in which of the 25 Munich city districts from Altstadt-Lehel to Laim the person under consideration lives
City district part	Indicates in which of the 108 Munich city districts parts the person under consideration lives
City district quarter	Indicates in which of the 448 Munich city districts quarters the person under consideration lives
Street	Indicates the street of a person's address
House number	Indicates the house number of a person's address
Alpha suffix	Indicates a possibly existing alpha suffix of the address of a person
Age	Indicates the age of a person in years
Gender	May contain the expressions "male" and "female"
Religious affiliation	Religious affiliation with the attributes "roman catholic", "evangelical", "other" and "none".
Marital status	Current marital: "mature/single", "married/in registered civil partnership", "divorced/dissolved civil partnership", "widowed/civil partnership dissolved by death" and "unknown"
Migration background	Binary variable, whether a migration background is present or not
Duration of residence in Munich	Indicates the duration of residence of a person in Munich in years
Immigration area	Informs about previous place of residence with the attributes "born in Munich", "Germany", "EU foreign countries", "Non-EU foreign countries" and "unknown"
Immigration area in Germany	Informs about previous place of residence with the attributes "born in Munich", "Region 14", "Upper Bavaria (without region 14)", "Bavaria (without Upper Bavaria)", "Germany (without Bavaria)", "Foreign countries" and "unknown"
Immigration	Informs about previous place of residence with the attributes "born in Munich", "Region 14", "Upper Bavaria (without region 14)", "Bavaria (without Upper Bavaria)", "Germany (without Bavaria)", "EU foreign countries", "Non-EU foreign countries" and "unknown"
Age at immigration	Indicates age of a person when moving to Munich

Table 7.1: Description of the Munich population data set

With the two newly added variables, the data set for the practical analysis of the small area estimation now comprises 17 variables for the 1,578,051 main residents of Munich.

As a further data source, geodata from the GeodatenService of the City of Munich were used. These are shape-files of the different administrative spatial levels of the city, which are also taken into account in the population data, namely the city districts, the city district parts and the city district quarters. These data are needed for the spatial presentation of the results.

Now that the data basis for the practical part of this thesis has been introduced, the central results of the descriptive analysis for the most important variables are presented below.

7.1.3 Descriptive analysis of the central variables of the Munich population data

This section starts with the duration of residence in Munich, as this characteristic will be the target variable in the context of the small area estimation of the population data. Therefore, it is important to get a more accurate understanding of this characteristic. The duration of residence of Munich's main resident population ranges from zero to 110 years in April 2022. The duration of residence is not available for only 218 observations in the dataset. This corresponds to only about 0.01 per cent of Munich residents. The arithmetic mean is 19.46 years, whereas the median is exactly 13 years. This is a first indication of a strongly right-skewed distribution of the duration of residence, which can also be clearly seen in the histogram in figure 7.1.

The green bars each cover a period of residence of five years, which means that the first bar shows that in April 2022 almost every third person in Munich will have lived in Munich for less than five years. Between five and 45 years of residence, the number of observations always decreases in direct proportion, with the 75 percent quantile being reached at 30 years. From a period of residence of 45 years onwards, the number of observations always decreases slightly every five years, before almost no cases can be detected from a duration of residence of 90 years onwards. The solid red line represents the arithmetic mean, whereas the blue dashed line represents the median length of residence in April 2022.

The spatial levels of the dataset

In the context of small area estimation, it is also relevant to descriptively analyse the duration of residence in the individual areas before estimation, as this is to be estimated later in the small area estimation. First, the areas themselves are considered. The dataset contains three administrative levels of local breakdown, namely the city districts, the city district parts and the city district quarters. The first thing to mention is that no clear spatial pattern is

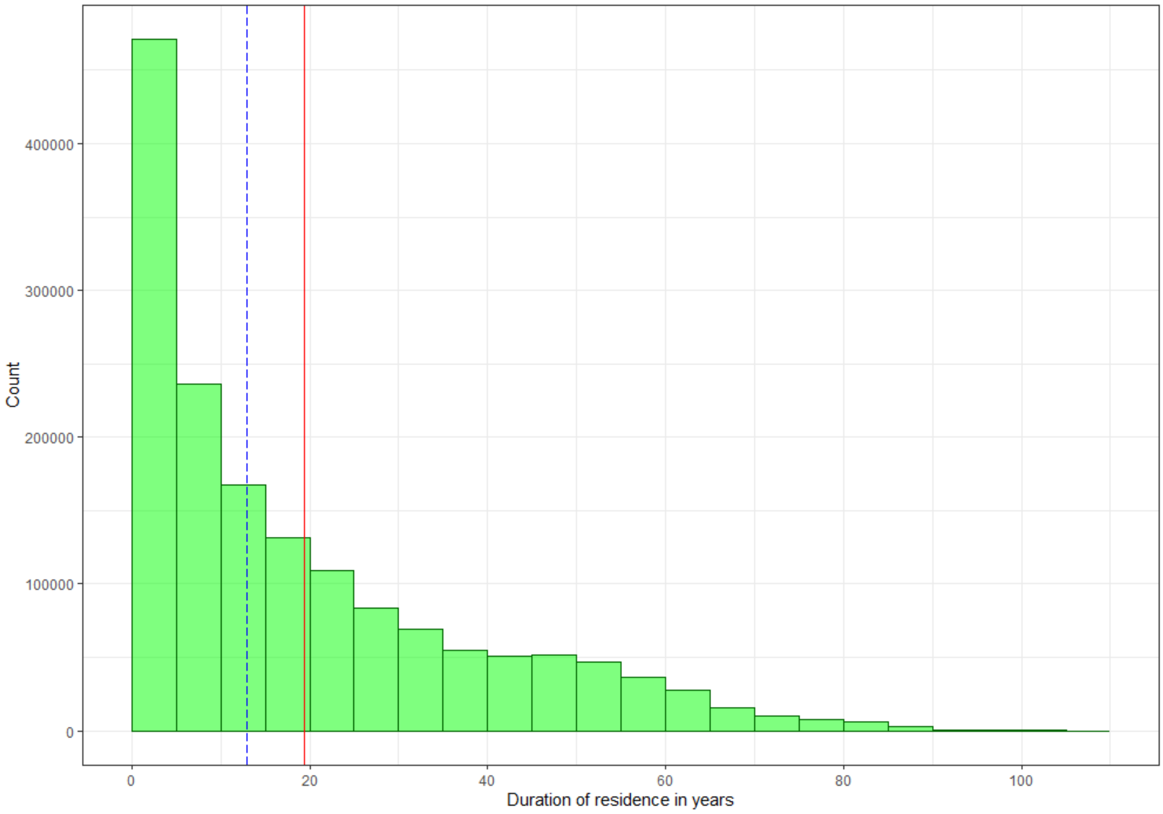


Figure 7.1: Histogram of the duration of residence of Munich's main resident population in April 2022

apparent in the 218 missing values for duration of residence. Every district has at least one resident with a missing value for this variable, and there is no district with more than 20 missing values for the duration of residence. The maximum value at the district part level is 10 and at the quarter level the maximum is 6. This makes it clear that there is no area in which all the missing values for this characteristic are concentrated.

Next, the sizes of the areas at the different levels are considered. The almost 1.6 million people with their main place of residence in Munich in April 2022 are spread across 25 city districts, with population figures in the areas ranging from around 20,000 in the district Altstadt-Lehel to just under 120,000 in Ramersdorf-Perlach. The average number of inhabitants at city district level is around 63,000. In 16 of the 25 city districts, the population in April 2022 is between 50,000 and 80,000 and thus only deviates from the mean by a maximum of around a quarter. Only three districts have more than 100,000 inhabitants. And even the smallest district still has a large number of inhabitants, so that an estimation with direct estimators should not be problematic.

The next step is to look at the sizes of the areas at city district part level. In the individual 108 city district parts, the number of inhabitants ranges from 149 to 51,267. The mean value is just under 15,000 inhabitants and the median of around 12,100 people indicates that half of the city district parts have fewer than 12,100 residents and the other half have more.

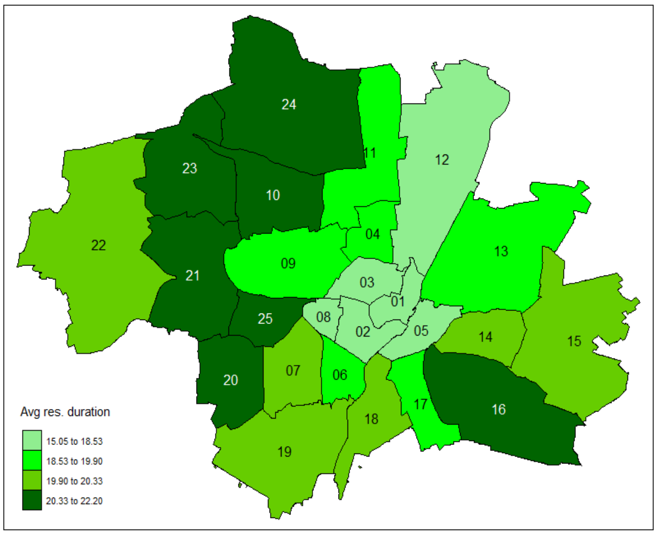
While on a spatial level above, the city district with the fewest inhabitants still had a size of more than 20,000 residents, the minimum is now much smaller with less than 150 residents. Here, estimation with direct estimators could already lead to imprecise results. In the initial application of their estimator, Fay and Herriot (1979: p. 272 ff.), when considering small places, primarily examined states with fewer than 500 and fewer than 1,000 inhabitants. At the city district part level, this number is still within limits. Only the part with the minimum value of 149 inhabitants has fewer than 500 residents, and only two other city district parts have fewer than 1,000 inhabitants.

The number of these small areas becomes considerably larger one level below. For the 448 city districts quarters in the data set, the number of inhabitants ranges from 36 to 12,695. It can therefore be seen that the largest quarter has far fewer inhabitants than the smallest district. The mean number of residents in the city district quarters is around 3,500 and the median is just under 3,000. In a total of 54 quarters, the number of inhabitants is below 500, which corresponds to a share of around 12 percent. And in 91 city district quarters the number of residents is less than 1,000, which is already more than 20 per cent of all quarters. This must definitely be taken into account in the context of small area estimation. But next, the average duration of residence in the individual levels is visualised graphically in order to illustrate the utility of small area estimation on the current data set once again.

The average duration of residence in the different spatial levels

For this purpose, it is started with the arithmetic mean of the duration of residence at city district level, which is shown in figure 7.2. Here the different city districts are coloured according to the quartile of the mean duration of residence. This indicates that the 25 percent of districts with the lowest mean duration of residence are coloured the lightest green, and the longer the duration of residence, the darker the colour. The city districts coloured in dark green represent the 25 percent with the highest mean period of residence in April 2022.

Overall, the average duration of residence in the districts is between approximately 15 and 22.2 years. Of these, the five city districts directly in the city centre and Schwabing-Freimann, which borders it to the north, have the lowest average durations of residence in April 2022. The city districts with a mean duration of residence in the second quartile are mostly close to the city centre and border at least one city district with an average duration of residence in the first quartile. The third quartile of durations of residence extends mainly across the south of the city, from Aubing-Lochhausen-Langwied in the west to Trudering-Riem in the east. With the exception of the south-eastern district of Ramersdorf-Perlach, the city districts with the highest average durations of residence are all concentrated in the west of Munich and border on each other. Overall, it can be stated that the high mean durations of residence are more likely to be found in the east and west of the city, whereas the average duration of residence is noticeably lower in the city centre and the districts close to the city centre.



1	Altstadt – Lehel
2	Ludwigsvorstadt – Isarvorstadt
3	Maxvorstadt
4	Schwabing West
5	Au – Haidhausen
6	Sendling
7	Sendling – Westpark
8	Schwanthalerhöhe
9	Neuhausen – Nymphenburg
10	Moosach
11	Milbertshofen – Am Hart
12	Schwabing – Freimann
13	Bogenhausen
14	Berg am Laim
15	Trudering – Riem
16	Ramersdorf – Perlach
17	Obergiesing – Fasangarten
18	Untergiesing – Harlaching
19	Thalkirchen – Obersendling – Forstenried – Fürstenried – Solln
20	Hadern
21	Pasing – Obermenzing
22	Aubing – Lochhausen – Langwied
23	Allach – Untermenzing
24	Feldmoching – Hasenbergl
25	Laim

Figure 7.2: Arithmetic mean of the residential duration in the 25 city districts of Munich in April 2022

However, as it has already been noted in this chapter that most city districts contain at least 50,000 inhabitants, the average durations of residence are an aggregation of many single individuals. To get a more accurate picture of the spatial distribution of durations of residence, it is also worth looking at levels below the city district level. This is done in figure 7.3, where the upper graph analyses the mean durations of residence at the city district part level and the lower graph looks at the arithmetic mean of durations of residence in April 2022 at the city district quarter level.

The first difference between the two graphs in 7.3 and also to the graph 7.2 can be seen in the legend. Since in all three cases the colours stand for the quartiles of the average duration of residence, the limits between the graphs of the spatial levels differ and each colour always stands for various values. Furthermore, one sees that the deeper one goes, the larger the range of the mean durations of residence becomes. At the city district part level, it is around eight to 24 years. At this level, there are also relatively few differences to the situation at the city district level. The parts with the low average residential durations in April are clustered near the centre and close to it. And as in figure 7.2, the long-term residents are predominantly to be found in the west and east of Munich. However, there are also some important differences between the city district part level and the city district level. For example, in the very west and very east of the city, there is a part whose mean duration of residence is in the lowest quartile, while the entire city district in which the part is located still has a duration of residence that is above the median. On the other hand, there are also city district parts close to the city centre that tend to have longer residential durations, and in the district of Schwabing-Freimann there are even parts in the quartile with the longest residential durations, although the entire district is in the lowest quartile of average residential durations. Thus, one can already see a certain gain in information by carrying out the spatial analysis

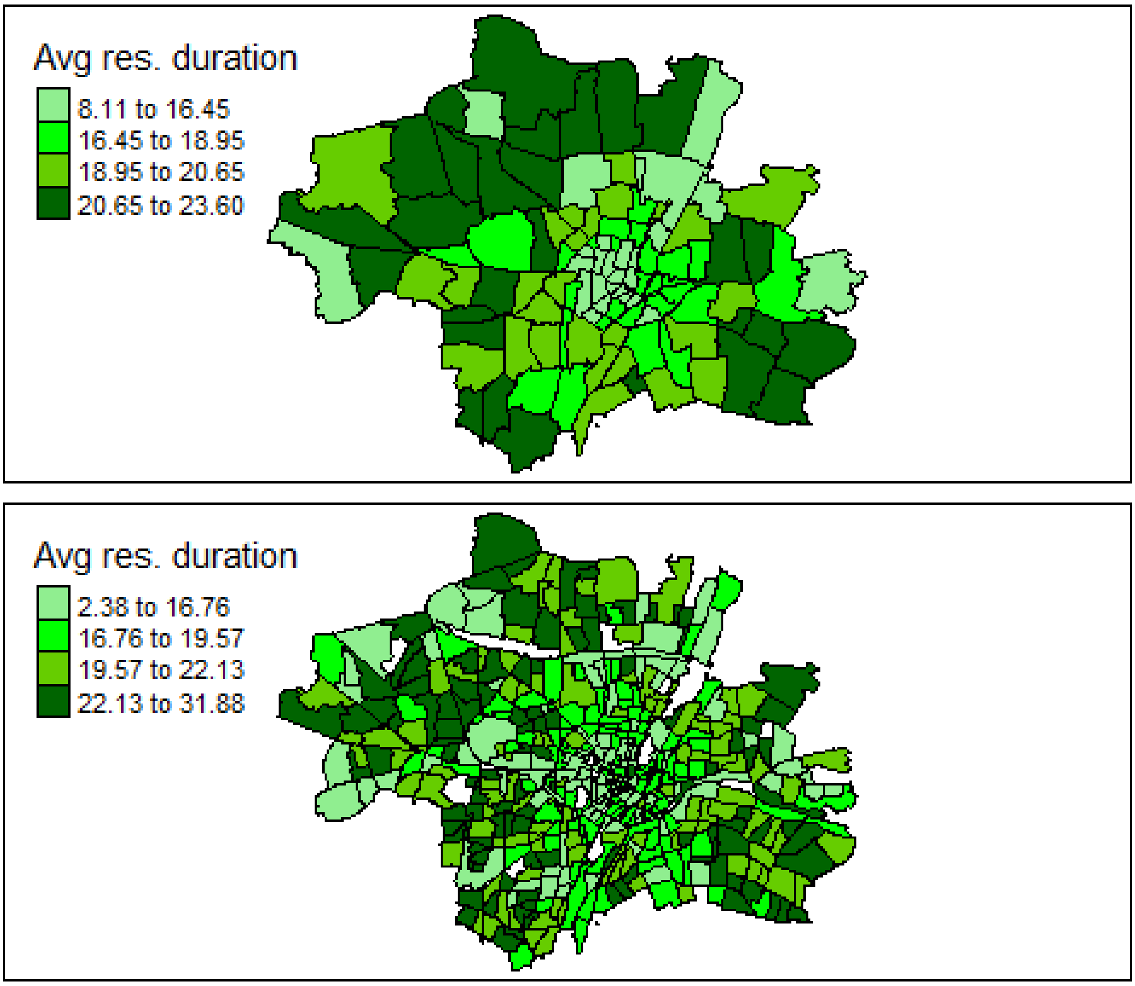


Figure 7.3: Arithmetic mean of the residential duration in the 108 city districts parts (top) and the 448 city district quarters (bottom) of Munich in April 2022

one level deeper. An even more informative view can be obtained by looking at the lower graph with the city district quarters in figure 7.3.

Instead of 105 city district parts, 468 quarters are now being examined. The quarters without colour in white are those for which the statistical office removed the values before the data transfer due to data confidentiality reasons. Overall, the average duration of residence at this level is between approximately two and 32 years. This range is four times as large as at city district level. Again, one can see the very low mean durations of residence in April 2022 especially in the city centre. However, there are quarters in all directions whose average duration of residence is in the first quartile. Also, the very large residential durations are now spread all over the city. But clusters of them can be seen in the north-west and south-east of Munich. In the quarters where the average duration of residence is in the second and third quartiles, the situation is no longer so evident. These are relatively uniformly distributed across the entire city. This deeper geographical breakdown provides information that would not have been available if the data had only been taken at the city district level. This helps in urban planning as a more accurate picture is now available even for areas with few inhabitants.

One aspect is still interesting to mention here, namely the correlation between the number of inhabitants in an area and the average duration of residence. For this purpose, the correlation coefficient according to Bravais-Pearson is considered at the three levels examined. This is around 0.215 at city district level. At the city district part level the coefficient is 0.488 and at the city district quarter level it is only 0.091. It can thus be seen that at the quarter level, where an area contains considerably fewer inhabitants on average than at the other two levels, there is almost no longer any correlation between the number of inhabitants and the average duration of residence in this area. Before that, a much more positive correlation between the two characteristics could be seen, especially at the city district part level. Now that the first descriptive analyses have been carried out for the small areas, the various estimators will be applied to these areas.

7.2 Small area estimators for the Munich residential durations

The presentation of the small area estimators for the Munich population dataset proceeds in several steps. First, the results for a large number of the estimators presented in this thesis are examined on the unmodified original data set, before the behaviour of the small area methods in the context of deficient data on modified data is examined in further steps.

7.2.1 Small area estimates for the original dataset

Since the unprocessed Munich population data from April 2022 is a complete survey, the small area estimation first requires samples of the whole data. For this, the same approach is used for all of the estimators: A ten percent sample is taken from each of the city district quarters. This results in samples ranging in size from 3 to almost 1300 observations for the individual quarters. No equally large sample sizes are consciously chosen here. On the one hand, this is because the size of the different quarters deviates too much from each other and, on the other hand, the small sample sizes are of particular relevance for small area estimation. Thus, with a total of 91 city district quarters, more than one fifth of them have a sample size of less than 100 observations and 54 quarters have a sample size of less than 50 observations. The respective estimators and the associated variances are then calculated on these samples. To avoid sampling effects, this procedure is carried out ten times on different samples for each of the small area estimators and the calculated results of the ten runs are averaged. This procedure can be seen as a kind of bootstrapping, as ten different samples are drawn from the same population for the analysis. It is quite possible that an element from the population does not appear in any or all of the samples. Due to the fact that in all city district quarters, in addition to the sample, the variable of interest is also available for all observations, the results of this procedure can be perfectly compared with the true data.

The analysis is started with the Horvitz-Thompson estimator. The results of this design-based direct estimator are presented in figure 7.4. The top left part again shows the arithmetic mean duration of residence in the 448 Munich city district quarters in April 2022 as the bottom image in figure 7.3. At the top right are the mean durations of residence for the individual quarters based on the estimates of the Horvitz-Thompson estimator. The part at the bottom left of figure 7.4 illustrates the differences between the estimated mean durations of residence and the true values for all city district quarters. And on the bottom right, the figure is completed by a plot of the variances of the estimation with the Horvitz-Thompson estimator. The more detailed representations and colours in the graphs are explained in the following.

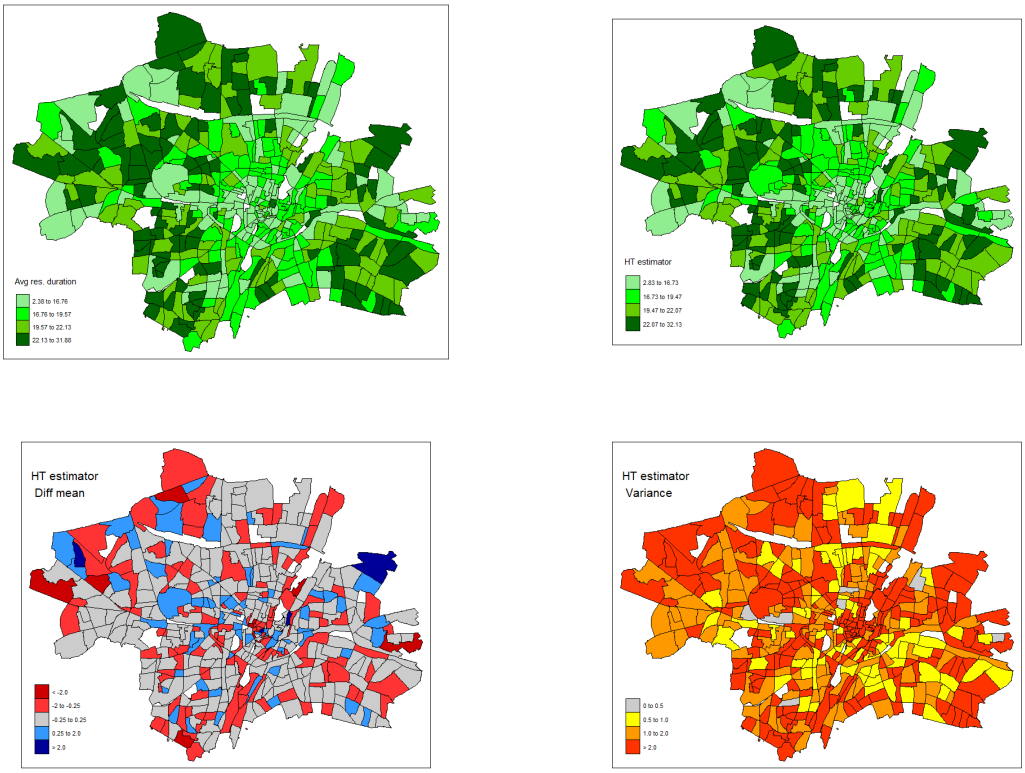


Figure 7.4: Results of the Horvitz-Thompson estimator for the original Munich population data
Top left: True means of residential duration in the 448 city district quarters
Top right: Estimated means of residential duration for the city district quarters
Bottom left: Differences between true and estimated means
Bottom right: Variances of the Horvitz-Thompson estimates for the quarters

Looking only at the upper part of the graph, one sees at first glance only a few differences between the true values and the estimates. For both graphs, the same colour scheme as in the previous section is used and both graphs are coloured according to the quartiles of the mean duration of residence. The legend indicates that the boundaries of the quartiles differ only minimally from each other. As with the true data, the high estimated durations of residence according to the Horvitz-Thompson estimator tend to be found in the northwest and southeast of the city. The low durations of residence are still predominantly clustered in the city centre. To better perceive the differences, it is worth taking a look at the map at the bottom left of figure 7.4. The difference in the mean duration of residence between the estimates and the true value is shown here. For this purpose, the true mean values of the duration of residence based on the whole data set were subtracted from the arithmetic mean of the ten estimates with the respective ten percent sample for each city district quarter. The white areas are still the quarters that were removed for data protection reasons. In the grey areas, the difference between the estimate and the true value lies in the interval between -0.25 and 0.25. The red areas represent an underestimate of the true value. The darker the red, the smaller the Horvitz-Thompson estimator is compared to the true value. For the light red areas the underestimation is between 0.25 and 2.0 and for the dark red areas it is even larger. The blue areas represent an overestimation of the true value. Here the Horvitz-Thompson estimator is larger than the true value. In the light blue area it is between 0.25 and 2.0 and in the dark blue area it is even larger. This scale is used for this type of representation for all other estimators.

For the Horvitz-Thompson estimator one recognises relatively many grey areas. This is not a big surprise, since for the Horvitz-Thompson estimator as an expectation-true estimator no large deviations between the estimator and the true value are expected. Otherwise, there are clearly more red than blue areas, which suggests that the Horvith-Thompson estimator tends to underestimate the mean duration of residence in April 2022. Summing up the deviations of the 448 quarters, one arrives at a value of just under -50. If one takes the absolute values for the summation, the result is around 165. From this one can conclude that the average deviation per quarter is below 0.4. Later one will see that this is a very low value compared to the other estimators. On the other hand, in the map to the right in figure 7.4 one can recognise that the variance of the Horvitz-Thompson estimator is very large. This graph shows the variances of the estimates for each area. The darker the colour, the greater the variance. In the grey city district quarters it is less than 0.5, but this is only the case in 1.5 percent of the quarters. In the yellow areas, the variance is between 0.5 and 1.0, which is the case in 76 areas. Light orange stands for a variance between 1.0 and 2.0 and in the dark orange quarters the variance of the estimator is even greater. This colour scheme for the variance is also used for the other estimators. In the Horvitz-Thompson estimator, more than half of the city district quarters are coloured dark orange. While the average variance of the Horvitz-Thompson estimator is just under nine in all quarters, the average for the 54 city district quarters with a sample size of less than 50 is over 52.

Next, the basic synthetic estimator is considered, which is referred to as BSE in the maps in figure 7.5. In addition to the basic synthetic estimator, this figure 7.5 also shows the difference between the true mean and the estimated value as well as the variance of the estimator for the post-stratified estimator, the sample size dependent estimator, the generalised regression estimator and the area specific generalised regression estimator. In the basic synthetic estimator and the post-stratified estimator, the population is further subdivided into strata or subgroups in addition to the subdivision into areas, since the group sizes serve as auxiliary information for estimation, as already explained in the sections on these two estimators. In the present case of the Munich population data, age and gender are used as grouping characteristics. The limits of the age groups are set at 18, 35 and 65 years, so that four groups can always be formed for both sexes.

It can already be seen at first glance that for the basic synthetic estimator, both for the difference between the true value and the estimated value and for the variance, the most city district quarters are coloured among all estimators in figure 7.5. Thereby, the number of quarters in which the true value is overestimated by BSE and the quarters in which it is underestimated is quite balanced. For both colours, one can also see a large number of dark-coloured city district quarters with a large deviation from the true mean. Of all 448 city district quarters, only 25 are coloured grey, which means a small difference between estimate and true value. In total, the difference is around 170 when added up over all quarters and just under 1,000 when the absolute amounts are used. In both cases, this is a much larger value than for the Horvitz-Thompson estimator, which was also to be expected since the basic synthetic estimator is a biased estimator. What is surprising, however, is that when looking at the variance of the BSE estimator, considerably more areas are coloured dark orange than in the Horvitz-Thompson estimator. Although the average variance of the BSE in the city district quarters is only half that of the Horvitz-Thompson estimator at around 4.5, it is greater than two in over 70 per cent of all quarters and thus a good 80 city district quarters are more dark orange in colour with the BSE than with the Horvitz-Thompson estimator. On the other hand, the basic synthetic estimator is the only one of all the estimators examined which, on average, has a lower variance in the small quarters with a sample size of less than 50 observations than in the remaining quarters. The explanation for the high deviations and the high variances in BSE could be that the duration of residence as the variable of interest does not have nearly the same mean value in the different groups. Due to the high correlation of age with duration of residence, the latter is significantly greater on average in the older age groups. In addition, the mean age varies greatly between the different city district quarters. Thus, the criterion for an approximatively unbiased estimator demanded in the section on the basic synthetic estimator is not satisfied. Even if one uses the age at move-in for the analysis instead of the age, which is significantly less strongly correlated with the duration of residence, the desired characteristics do not improve. Instead, the post-stratified estimator, which is considered next, can help in such circumstances.

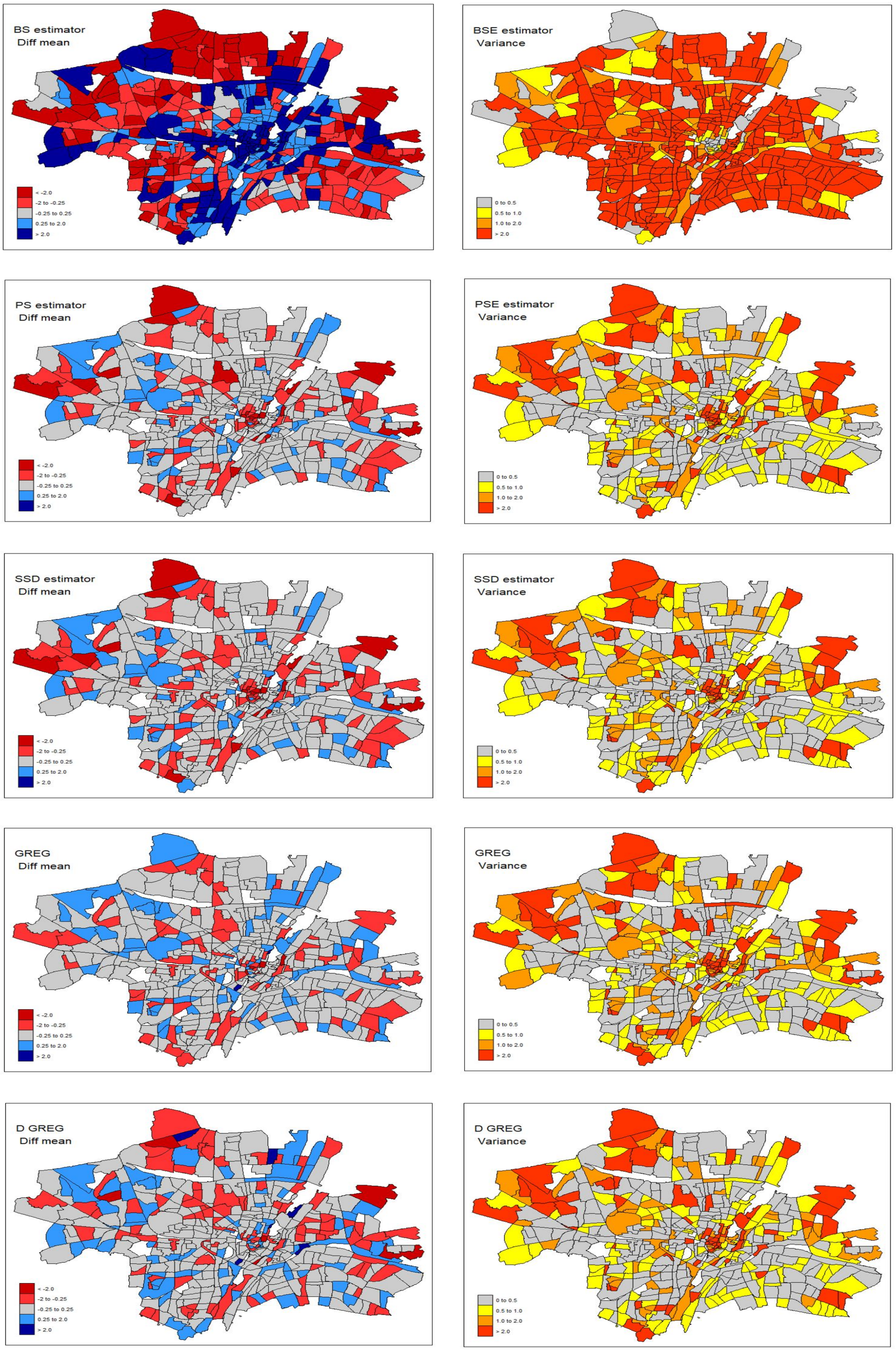


Figure 7.5: Difference between estimates and true means (left) and variances of estimates (right) for further estimators
First row: Basic synthetic estimator; Second row: Post-stratified estimator;
Third row: Sample size dependent estimator; Fourth row: Generalized reg. estimator;
Fifth row: Area based generalized reg. estimator

This estimator is shown in the second row in figure 7.5 and one can immediately see that it has considerably fewer coloured areas than the basic synthetic estimator. For over 60 percent of all areas, the amount of deviation of the estimated mean from the true mean is less than 0.25, reflecting the large number of grey areas. The summed amount of the absolute deviations across all 448 city district quarters is, with a value of around 250, only a quarter of the value from the basic synthetic estimator. The average variance of about 1.15 per area is also only about a quarter as large as for the basic synthetic estimator, and the map shows that far fewer areas are coloured dark orange for the variance of the post-stratified estimator. As was already evident from the Horvitz-Thomspon estimator, the variance of the post-stratified estimator in the areas with a sample size of less than 50 is clearly larger than for the other areas. Overall, the post-stratified estimator can be considered a better estimator than the basic synthetic estimator, since both the deviations from the true value and the variances of the estimates for the individual areas are notably lower for the post-stratified estimator.

The sample size dependent estimator serves to combine the advantages of the two estimators just discussed. This is usually done by specifying a value for δ. Typical values here are $\frac{2}{3}$ or 1. In the present case, δ was set to the value 0.99. With a value of 1, γ would also take the value 1 due to the sample design and only the information of the post-stratified estimator would be used. The value 0.99, on the other hand, means that components of the basic synthetic estimator are also included in the estimation. On the two maps in the third row of figure 7.5 there are no major differences to the two maps of the post-stratified estimator. Even though the differences are small, the sum of squared deviations between estimated and true value has decreased by more than five units and also the variance per city district quarter has declined by more than 0.1. Thus, at least a small improvement can be observed by using the sample size dependent estimator, which would be even greater if the basic synthetic estimator performed better.

Next, a model-assisted estimator is considered with the generalised regression estimator. A distinction is made between $\hat{Y}_d^{GREG}$ from formula 4.6, where all information is used to estimate β, and $\hat{Y}_d^{GREG^*}$ from formula 4.7, where the estimation of β_d is area specific. For the estimator $\hat{Y}_d^{GREG}$ as an approximate unbiased estimator, a smaller deviation between the estimate and the true values is expected in theory, but a larger variance than for $\hat{Y}_d^{GREG^*}$. When estimating the models for the two generalised regression estimators, age, gender and migration background are used as independent variables. The results of these two estimators are presented in the last two rows in figure 7.5, where the area specific estimator $\hat{Y}_d^{GREG^*}$ is referred to as D GREG in the maps and is found in the bottom row. For the deviations from the true mean, about one third of the areas are coloured in both cases. This means that in many areas the deviation of the estimate from the true value is only very small. Nevertheless, there are also differences in the use of these two estimators, which can be seen very well in the northern city district quarters, for example. There, the area-specific generalised regression estimator underestimates the mean duration of residence in the many red quarters. With the estimator $\hat{Y}_d^{GREG}$, on the other hand, the proportion between red, blue and grey areas in the north of Munich is quite balanced. For the area-specific version of the generalised regression estimator, the summed absolute deviation between the true value and the estimated mean across all areas is just under 164, and for $\hat{Y}_d^{GREG}$ it is even smaller at around 138. Thus, the generalised regression estimator has the lowest deviations among all the estimators presented so far. In the two maps with the variances of the estimators of the generalised regression estimator, one sees at first glance only very few differences between the two estimators. However, if one looks at the corresponding numbers, large differences become apparent. The figures of the deviations and variances of all estimators presented so far can also be found in table 7.2. For the estimator $\hat{Y}_d^{GREG}$, the average variance in the individual quarters is just under three and in the small areas with sample sizes of less than 50 it is over 18. The same ratio can be seen for the area specific version of the estimator, although at a much lower level. Overall, the variance is less than 1.5 per quarter and in the small areas it is around 6. Thus, the generalised regression estimator meets the theoretical expectation that the area-specific version shows larger deviations from the true mean, but that the variance is noticeably lower.

Estimator $\hat{\bar{Y}}$	**Sum of deviations**	**Sum of absolute deviations**	**Variance per area**	**Variance per area in the small areas**
Horvitz-Thompson	- 49.16	166.93	8.80	52.72
Basic synthetic	169.26	993.39	4.32	0.35
Post stratified	- 172.24	256.11	1.16	4.20
Sample size dependent	- 164.78	250.76	1.14	4.08
Generalised regression	- 11.24	138.54	2.89	18.48
Area spec. GREG	- 5.60	163.95	1.34	6.13

Table 7.2: Deviations from the true mean and variances of the previous small area estimators

Table 7.2 also serves as a good summary of the results so far. One can clearly see the large difference between the variance of the Horvitz-Thompson estimator and the variance of the other estimators, especially for the small areas. This highlights the problem that when using direct estimators, the estimates are no longer precise, especially in areas with small sample sizes. The second largest variances in the small areas occur in the generalised regression estimators, which are also classified as direct estimators in this thesis. Here, the difference to the other estimators is not so large in relation to all areas, but only becomes visible with this clarity in the city district quarters with a sample size of less than 50. After the detailed presentation of the practical results of these estimators from the fourth chapter, the next step is to analyse the behaviour of the Battese-Harter-Fuller estimator and the Fay-Herriot estimator from the fifth chapter for the Munich population data.

A total of three estimators are discussed here. In addition to the Battese-Harter-Fuller estimator and the classical Fay-Herriot estimator, a spatial Fay-Herriot estimator, which was presented in the section 5.6 on other model-based small area methods, is also considered. In all three cases, age, gender and migration background were again used as covariates for the underlying model. The results of the practical analysis of these three estimators are visualised in figure 7.6.

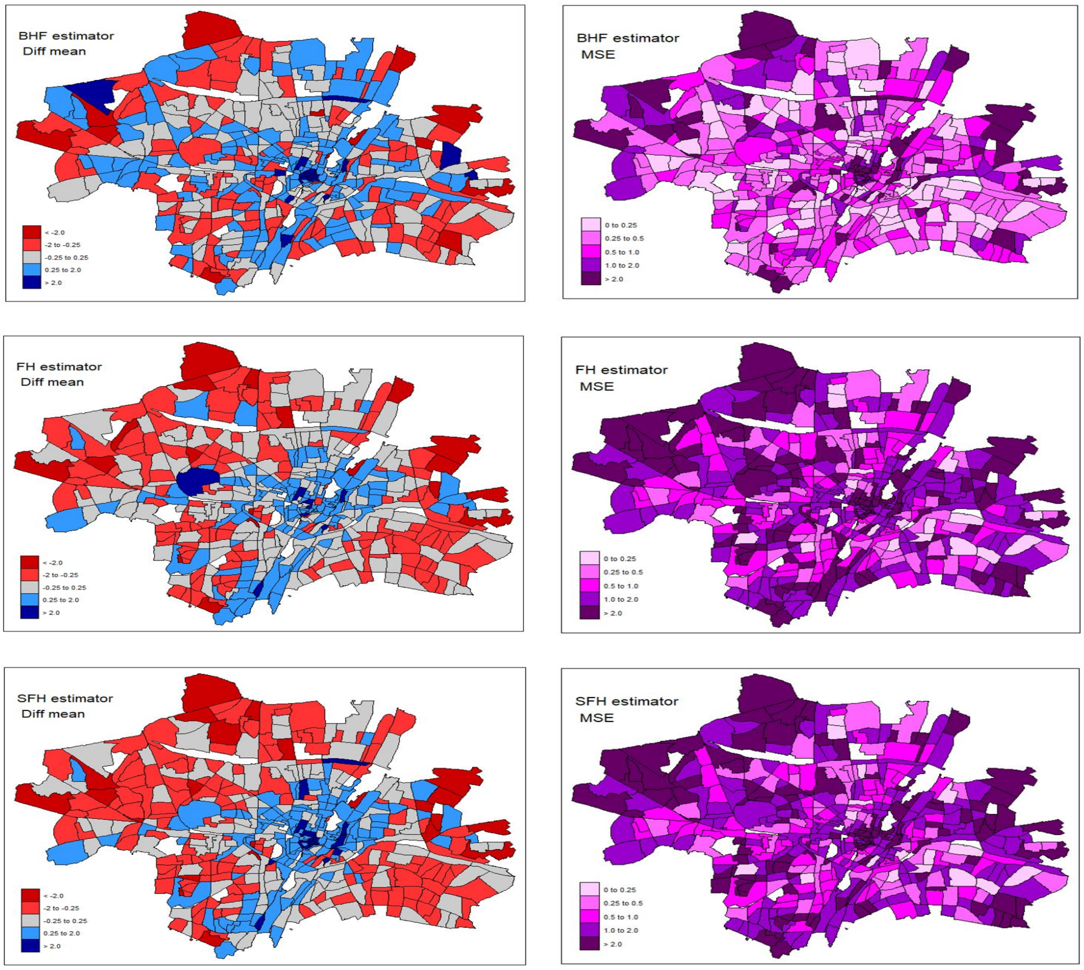


Figure 7.6: Difference between estimates and true means (left) and mean squared errors of estimates (right) of the model based small area estimators
First row: Battese-Harter-Fuller estimator
Second row: Fay-Herriot estimator
Third row: Spatial Fay-Herriot estimator

In this figure, the deviations of the estimated values from the true mean value of the entire data are again shown on the left side. The maps on the right side represent the mean squared error (MSE) of the estimators. By using the mixed model with the random effects, the variance estimation of these three small area methods no longer corresponds to the procedure for the previously presented estimators without random effects. Instead, the R-package *sae* (Molina, Marhuenda 2015: p. 81 ff.) provides functions that calculate the MSE of the estimates based on bootstrap iterations. The MSE for the estimator of an area is larger the darker the colour in the maps in figure 7.6. The breaks are set to 0.25, 0.5, 1 and 2.

First, the results of the Battese-Harter-Fuller estimator are considered. It tends to overestimate the true value, as can be seen from the many blue areas on the map. Therefore, the number of summed deviations from the true mean value is just over 50. The sum of the absolute deviations is just under 330, which is larger than most of the estimators presented so far. However, this is due to the fact that the BHF, as a composite estimator, consists, among other things, of typically biased synthetic estimators. At the same time, it can be seen that only very few city district quarters have a very large MSE in the estimation, as only very few areas are coloured dark purple. The average mean squared error per area is only 1. In the 54 small city district quarters, the average value is just over 4.

The situation is different for the two area level estimators. In these situations, the high number of red areas tends to underestimate the true average duration of residence. In both models, the summed number of deviations across all areas also takes on a negative value of -54 for the Fay-Herriot estimator and -42 for the spatial version. If one adds up the absolute amounts, one obtains values of 285 and 357 respectively. This shows that for the Fay-Herriot estimator the deviations are overall smaller than for the Battese-Harter-Fuller estimator, whereas they are larger for the spatial Fay-Herriot estimator. With the spatial Fay-Herriot estimator, only the neighbourhoods of the individual city district quarters are considered. The estimation uses a proximity matrix that takes the value 1 for city district quarters with a common border and the value 0 if two quarters are not bordering. If one looks at the mean squared error of the two area level estimators, one does not see any major differences in the maps between the two estimators, but one does see clearly more dark-coloured areas for both than for the Battese-Harter-Fuller estimator. This is also expressed in the numbers, which are summarised again for the three model-based estimators in table 7.3. So, the average MSE per area for both the Fay-Herriot and the spatial Fay-Herriot estimator is around 1.5, which is about one and a half times larger than for the unit level model. In the small areas with a sample size of less than 50 observations, on the other hand, the average mean squared error is about 3 and is thus smaller than for the Battese-Harter-Fuller estimator.

Estimator $\hat{\bar{Y}}$	**Sum of deviations**	**Sum of absolute deviations**	**MSE per area**	**MSE per area in the small areas**
BHF	51.44	328.17	1.06	4.35
FH	- 54.30	285.80	1.54	2.90
Spatial FH	- 42.12	357.06	1.55	3.02

Table 7.3: Deviations from the true mean and mean squared errors of the model-based small area estimators

It can therefore be summarised that the use of spatial information does not bring any noticeable improvement compared to the Fay-Herriot estimator. However, no clear conclusion can be drawn between the Fay-Herriot estimator and the Battese-Harter-Fuller estimator. Overall, the deviations of the estimator from the true value are somewhat larger with the BHF estimator than with the FH, whereas the use of unit level information yields a lower mean squared error per area. But this does not apply to the small areas. Compared to the six estimators presented in chapter four, the model-based estimators show a considerable reduction in variance, especially in the small areas, as the MSE is composed of the bias and the variance.

Now that the results of nine estimators have been analysed and compared with each other on the original data of the Munich residential durations of April 2022, the behaviour of selected estimators with deficient data will be examined in the further course of this chapter.

7.2.2 Small area estimates for different kinds of deficient data

This section examines the behaviour of certain estimators with respect to deficient data. For this purpose, the Munich population dataset is artificially modified to obtain different types of deficit data. The exact procedure is described at the respective position.

Small area estimation in the context of outliers

In the descriptive analysis of the duration of residence, a strongly right-skewed distribution was found. In addition to many very short durations of residence, there are also some outliers with a duration of residence of up to 110 years. In this part of the thesis, it will be investigated whether the use of the median as a robust measure of location has an advantage over the arithmetic mean. It is assumed that especially in the 54 city district quarters with a sample size of less than 50 people, outliers have a particularly large influence. If these outliers enter the sample in one of the ten runs, it can be assumed that the estimator is biased upwards in this run. This run would also have a high influence when averaging the ten runs using the arithmetic mean. The median could help here, since with it outliers in the individual runs do not have that much weight. Therefore, in the following it is examined for the different estimators how the deviation of the estimated value from the true mean behaves when the median is used for averaging over the ten runs compared to the arithmetic mean. These results are shown in table 7.4.

Estimator $\hat{\hat{Y}}$	Sum of absolute deviations **Arithm. mean** all areas	Sum of absolute deviations **Arithm. mean** small areas	Sum of absolute deviations **Median** all areas	Sum of absolute deviations **Median** small areas
Horvitz-Thompson	166.93	66.42	198.22	75.94
Basic synthetic	993.39	227.18	993.05	227.00
Post stratified	256.11	174.69	296.76	201.43
Generalised regression	138.54	58.73	164.79	70.00
Area spec. GREG	163.95	60.49	197.97	78.69
Battese-Harter-Fuller	328.17	133.44	370.10	139.21
Fay-Herriot	285.80	93.63	300.79	100.90

Table 7.4: Deviations of the estimator averaged with the arithmetic mean and the median from the true mean in all areas and the small areas with a sample size of less than 50 people

The use of the median is shown in the two columns on the right, once summed up over all areas and once only over the 54 city district quarters with a sample size of less than 50. It can be seen that there are no clear improvements in bias for any of the estimators if the median is used instead of the arithmetic mean. Only the basic synthetic estimator shows minimally smaller deviations from the true value when averaging over the median, both over all areas and in the small areas. For some estimators, such as the post-stratified estimator or the Battese-Harter-Fuller estimator, one even sees considerably larger absolute deviations between the arithmetic mean of all data and the median of the ten estimates than between the arithmetic mean of the population data and the arithmetic mean of the estimators of the ten runs. At this point it should be explicitly mentioned again that the median considered is the median of ten runs of the estimation of an expected mean. Therefore, it makes no sense to compare the median of the ten different estimates of an area mean with the median of all data of the respective area. Much larger differences would arise here. The use of the median of the ten runs is only intended to counteract the case that the sampling of extreme outliers would extremely bias the estimate in individual runs, especially in small areas, but this does not seem to be the case on the basis of the results described.

One explanation for this could be the size of the city district quarters in which the outliers are located. The entire data set of almost 1.6 million observations includes only 13 people with a duration of residence of more than 100 years. All of these 13 people live in neighbourhoods with more than 1,500 inhabitants. In many cases, the number of inhabitants is even more than 5,000. As a result, the corresponding sample sizes are also far larger than, for example, in the city districts quarters with less than 500 inhabitants, since ten percent of the population is included in the sample everywhere. In the quarters with the large sample sizes, the outliers in the estimated variable do not have such an extreme effect on the estimate due to the many other observations.

Small area estimation with missing data and mean imputation

Next, the behaviour of some aspects of small area estimation will be considered in the context of missing data and imputed values. The population data set of the statistical office of the state capital Munich is a very high-quality data set in this respect. Of all the characteristics considered in this analysis, only the duration of residence has 218 missing values among all almost 1.6 million inhabitants. Therefore, for illustrative purposes, further missing values are artificially generated for duration of residence as the variable of interest.

For this purpose, different missing rates are used, which determine the number of na values in each of the ten runs in each city district quarter for this characteristic. The chosen missing rates mr are 1/10, 1/4, 1/3 and 1/2. The effects on the data set are described using the example of the missing rate $mr = 1/3$. Here, every third sampled value is declared na in every run in each quarter for the duration of residence. Thus, each of the ten sample data sets used now contains over 50,000 missing values in the duration of residence. This missingness can be considered MCAR, as the removal of the values was done completely randomly on every third sampled unit and was not made dependent on any variables. For the sampling design weights, instead of the previous value of $w_i = \frac{1}{0.1}$, the value $w_i = \frac{1}{0.1 \cdot (2/3)}$ is now chosen to take into account that for the duration of residence, of the ten percent of the sampled values, only a proportion of 2/3 is actually available. In general terms, the new weights based on the missing rate mr result in $w_i = \frac{1}{0.1 \cdot (1 - mr)}$.

To investigate the effects of imputation, a simple mean imputation is performed. For each missing value in the duration of residence, the arithmetic mean of this variable in the respective city district quarter is used. Subsequently, different estimators are calculated on the missing data set and the imputed data set, namely the Horvitz-Thompson estimator as a basic estimator, the post-stratified estimator because of its positive properties when analysed on the original data set, and the Battese-Harter-Fuller estimator and Fay-Herriot estimator as model-based estimators. First, the results for the two non-model-based estimators are considered. Similar to the analysis on the original data in table 7.2, these are presented in table 7.5 for all the missing rates used.

Estimator $\hat{\hat{Y}}$	**Data**	**Missing rate**	**Sum of deviations**	**Sum of absolute deviations**	**Variance per area**	**Variance per area in the small areas**
HT	Original	0	- 49.16	166.93	8.80	52.73
HT	Missing	1/10	- 135.97	365.89	9.01	51.99
HT	Missing	1/4	- 107.49	342.81	11.14	64.59
HT	Missing	1/3	- 151.50	355.04	12.31	71.11
HT	Missing	1/2	- 157.95	612.43	16.89	98.21
HT	Imputed	1/10	- 0.56	512.61	8.14	47.73
HT	Imputed	1/4	1.90	573.03	7.87	47.12
HT	Imputed	1/3	11.08	526.03	7.64	45.91
HT	Imputed	1/2	- 18.01	678.63	7.15	43.45
PSE	Original	0	- 172.24	256.11	1.16	4.20
PSE	Missing	1/10	- 210.60	389.44	1.22	4.05
PSE	Missing	1/4	- 223.29	411.55	1.38	4.24
PSE	Missing	1/3	- 280.79	448.83	1.45	3.90
PSE	Missing	1/2	- 352.48	672.60	1.77	3.90
PSE	Imputed	1/10	- 133.89	479.23	1.08	3.75
PSE	Imputed	1/4	- 127.65	535.12	0.96	3.16
PSE	Imputed	1/3	- 116.26	528.02	0.94	3.26
PSE	Imputed	1/2	- 130.77	683.68	0.69	2.02

Table 7.5: Deviations from the true mean and variances of HT estimator and PSE for missing data and imputed data compared to the original data for different missing rates

A similar pattern can be seen for both the Horvitz-Thompson estimator and the post-stratified estimator. In both cases, the sum of deviations of the estimators for the missing data for all missing rates is clearly more negative than for the original data, which means that the estimators for the missing data tend to underestimate the true values. For the imputed data, this value is the lowest in each case. On the other hand, for both estimators, the sum of the absolute values of the deviation between the estimator and the true value is the largest across all city district quarters for the imputations. With a higher missing rate, the absolute deviations tend to increase both for the data set with missing values and for the imputed data set with few exceptions. Especially in the case of the estimate with missing data, this sum increases enormously when the missing rate is increased from 0.10 to 0.50. This increase is not quite as high for imputed data. However, even with low missing rates, the sum of the absolute deviations between the estimated value and the true value is very large for the data with imputations. This may be due to the fact that in the imputed data set a lot of values have been inserted in the middle of the value range and the estimator is then also pulled more towards the middle of the data and then larger deviations from the true values at the edge of the distribution arise.

By inserting the many values in the middle of the value range, the estimator on the imputed data set has the smallest variance both across all areas and in the small areas with a sample size of less than 50 for all missing rates. This effect was to be expected, as it was already mentioned in the previous chapter that the variance tends to be underestimated in the imputed data sets. On the other hand, the average variance of the estimators in the city district quarters is considerably larger on the data set with missing data than on the original data. One explanation for this is that in the data set with the missing values, only a smaller part of the observations is available for estimation compared to the original data set. This effect becomes more evident the smaller the observed data set is. With the Horvitz-Thompson estimator, the variance increases considerably in all areas as well as in the small areas with a higher missing rate. With the post-stratified estimator, this effect is only seen when looking at the average variance across all areas. In the small areas, only minor changes in the variance of the estimators on the data set with the missing values are seen here with the modification of the missing rate. On the other hand, the variance situation for both estimators on the imputed data set is identical in all city district quarters as well as in those with a sample size of less than 50. Here, the variance decreases consistently with increasing missing rate, with one small exception for the post-stratified estimator. In comparison, the behaviour of the two model-based estimators on the missing and imputed data is examined in table 7.6.

For both the Battese-Harter-Fuller estimator and the Fay-Herriot estimator, the sum of the deviations between the estimated value and the true value on the imputed data is closest to zero for both estimators with all missing rates. On the other hand, this sum is largest for both estimators when looking at the absolute values of the deviations. For both model-based estimators, the absolute deviations for the data set with the missing values and the data with imputed values increase almost continuously as the missing rate increases. This situation is very similar to the two non-model-based estimators just considered.

Next, the mean squared error is considered for the model-based estimators. It should be noted here, that in addition to the variance, the deviations of the estimators from the expected value also play a role. This means that large deviations, as with the two estimators on the imputed data set, have a negative effect on the MSE. For the Fay-Herriot estimator,

Estimator $\hat{\bar{Y}}$	**Data**	**Missing rate**	**Sum of deviations**	**Sum of absolute deviations**	**MSE per area**	**MSE per area in the small areas**
BHF	Original	0	51.44	328.17	1.06	4.35
BHF	Missing	1/10	59.23	323.91	1.15	4.65
BHF	Missing	1/4	59.91	324.08	1.27	4.85
BHF	Missing	1/3	50.85	344.66	1.38	5.05
BHF	Missing	1/2	62.26	424.65	1.66	5.64
BHF	Imputed	1/10	47.52	356.46	1.07	4.37
BHF	Imputed	1/4	42.53	401.95	1.10	4.76
BHF	Imputed	1/3	41.80	403.48	1.08	4.75
BHF	Imputed	1/2	28.09	532.16	1.04	4.99
FH	Original	0	- 54.30	285.80	1.54	2.90
FH	Missing	1/10	- 87.18	382.47	1.65	3.01
FH	Missing	1/4	- 106.55	397.22	1.90	3.35
FH	Missing	1/3	- 132.26	402.73	2.08	3.64
FH	Missing	1/2	- 256.78	536.02	2.94	4.45
FH	Imputed	1/10	- 5.76	425.46	1.48	2.89
FH	Imputed	1/4	- 6.29	473.13	1.34	2.77
FH	Imputed	1/3	17.20	457.00	1.19	2.48
FH	Imputed	1/2	- 6.09	609.92	1.01	2.05

Table 7.6: Deviations from the true mean and mean squared errors of the model-based small area estimators in the context of missingness and imputed data with different missing rates

however, the MSE still behaves like the variance for the Horvitz-Thompson estimator and the post-stratified estimator. Here, the average MSE is lowest in all areas as well as in the small city district quarters on the imputed data and highest by far in the estimation on the data set with missing values. For the data with missing values, the mean squared error increases with a higher missing rate and for the estimation with imputed values, the MSE decreases in this case. This is true without expectation across all city district quarters as well as in the small ones with the sample size below 50. For the Battese-Harter-Fuller estimator, the mean squared error is also largest for the estimates with missing data. Here it also increases continuously with a higher missing rate. On the imputed data set, however, the MSE of the estimate is around the same as on the original data, and only with a high missing rate of 1/2, the MSE is smaller on the imputed data than on the original data in all areas. In the small areas, the MSE for all missing rates is not much lower for the imputed data than for the data with missing values. Now that these four estimators have been analysed in the case of missing and imputed data, a simulated data set with missing values and imputed data is analysed to compare the findings, before the effect of measurement error on the results of the four estimators in practice is looked at at the end of this chapter.

Small area estimation on a simulated data set with missing values and imputation

The goal of this section is to compare the results from the previous section and the behaviour of the small area estimators with the Munich population data to the results on a simulated data set. In summary, it can be repeated that all four estimators behave quite similarly. For all of them, the summed amounts of the deviation between the estimate and the true value tend to increase as the missing rate increases, both for the data with missing values and for the imputed data. The imputed data generally show a smaller variance or a smaller MSE than the original data, which decreases with increasing missing rate, while in the analysis with the missing data a larger value is observed for this parameter than in the original data, which increases with a higher missing share. Because of the similar behaviour, only the performance of the Horvitz-Thompson estimator on the simulated dataset is examined in this section. This dataset contains 300,000 observations and is divided into 273 areas. The areas have sizes of 50, 100, 200, 500, 1,000, 1,500, 2,000 and 2,500 observations respectively. This means that the behaviour of the estimators can also be compared on smaller areas, as was already the case with the Munich population data. Instead of the duration of residence in Munich, a variable is simulated as the variable of interest that corresponds to 300,000 times the draw from a normal distribution with mean 20 and standard deviation 4. The procedure for calculating the estimators works exactly as before. There are still ten runs. In each of them, a ten percent sample is drawn in each area on which the estimators are calculated. In addition, the same missing rates as in the previous chapter and the mean imputation are used to generate data with missing values and imputed data sets. The results of this are shown in table 7.7.

Much of the results of the analysis of the simulated data coincide with those on the Munich population data. In both cases, the sum of the deviations for all missing rates on the imputed data is clearly closer to 0 than on the data with missing values. Furthermore, the sum of the absolute deviations between the estimated value and the true value increases with increasing missing rate on both the data with missings and the imputed data, with exactly one exception. The behaviour of the variance of the estimators on the simulated data is identical to that with the population data. On all areas as well as on the 90 small areas of the simulated data with a sample size of less than 50, the variance on the data with missing values is always larger than on the complete data for all missing rates and on the imputed data it is always smaller. In addition, the variance decreases steadily with increasing missing rate in the imputed data and in the data set with missing values it increases steadily in this case.

Estimator $\hat{\bar{Y}}$	**Data**	**Missing rate**	**Sum of deviations**	**Sum of absolute deviations**	**Variance per area**	**Variance per area in the small areas**
HT	Simulated	0	8.97	46.42	16.69	44.05
HT	Missing	1/10	- 58.23	93.83	17.70	46.31
HT	Missing	1/4	- 165.30	181.57	20.25	52.48
HT	Missing	1/3	- 145.74	153.87	23.98	62.58
HT	Missing	1/2	- 117.03	155.68	31.58	81.90
HT	Imputed	1/10	8.70	48.75	16.60	43.80
HT	Imputed	1/4	9.66	55.55	16.58	43.76
HT	Imputed	1/3	8.54	58.14	16.54	43.67
HT	Imputed	1/2	3.21	63.91	16.36	43.17

Table 7.7: Deviations from the true mean and variances of the Horvitz-Thompson estimator for missing data and imputed data for different missing rates on the simulated dataset

Only a few small things differ between the estimation with the Horvitz-Thompson estimator on the population data and the simulated data. For example, the sum of the absolute deviations from the true value for the population data on the imputed data for all missing rates is considerably larger than for the data with the missing values. This is no longer the case with the simulated data. But otherwise there are no more major anomalies in the analysis with these two data sets. Therefore, in the next section the influence of measurement errors on the small area estimation for the Munich population data will be considered.

Small area estimation for data with measurement errors

Two different cases are considered in this section. First, the effects of measurement error for metric covariates are examined. This situation was also examined in the previous chapter in the measurement error bivariate Fay-Herriot model and in the approach of Polettini and Arima. Since, for example, the Horvitz-Thompson estimator does not use covariates, this case is only investigated for the two model-based estimators. In the second step, all four estimators are examined to see what happens if the dependent variable also has a measurement error.

For the model-based estimators, age, gender and the indicator whether there is a migration background are again used as covariates when estimating the duration of residence. For the first case (I), a random measurement error is added to the age. For this purpose, a value from a normal distribution with mean 0 and standard deviation 1 is drawn for each unit of the data set, which is added to the age of the respective observation. This measurement error is independent of the other variables as well as independent of the true value. The second row of each estimator in table 7.8 shows the effects of this measurement error in the age characteristic as one of the explanatory variables on the deviation of the estimate from the true value and on the MSE of the estimator. In the third row of each estimator in table 7.8, there are presented these effects for the second case (II) where, in addition to the measurement error in age, a measurement error in the duration of residence as the dependent variable was added. Again, for each respondent, a draw from a normal distribution with mean 0 and standard deviation 1 was added to the true value of duration of residence. Thus, again, the measurement errors for both variables are considered to be independent of the true value and independent of the other characteristics, since the addition of the measurement error for both variables is completely random and does not depend on other characteristics.

Estimator $\hat{\bar{Y}}$	**Sum of deviations**	**Sum of absolute deviations**	**MSE per area**	**MSE per area in in the small areas**
BHF **Original** data	51.44	328.17	1.06	4.35
(I) BHF with **measurement error** (age)	51.83	328.44	1.06	4.36
(II) BHF with **measurement error** (y + age)	52.50	330.78	1.06	4.37
FH **Original** data	- 54.30	285.80	1.54	2.90
(I) FH with **measurement error** (age)	- 41.72	392.52	1.57	2.97
(II) FH with **measurement error** (y + age)	- 41.41	394.06	1.57	2.97

Table 7.8: Deviations from the true mean and mean squared errors of the model-based small area estimators in the context of a measurement error for the covariate age and the duration of residence

With the Battese-Harter-Fuller estimator, in which the measurement errors are included in the estimate at unit level, almost no differences can be seen between the estimate with the original data and the estimation with the two cases of measurement errors. Neither adding the measurement error in age nor in duration of residence changes the deviations of the estimate from the true mean. This applies both to the summed deviations of all 448 city district quarters and to the sum of the absolute amounts of the deviations. The average mean squared error per area remains more or less identical in all three cases, and only in the case of the MSE in the small areas with less than 500 inhabitants, a really minimal increase in the mean squared error can be seen due to the addition of the measurement errors.

With regard to the MSE, the situation is identical for the Fay-Herriot estimator. Only a very small increase in the mean squared error can be detected by adding the measurement error across all areas as well as in the small areas compared to the estimation with the original data. In the case of deviations from the true value, on the other hand, the use of the Fay-Herriot estimator shows clearer distortions when measurement errors occur at area level. In this respect, it does not make a difference whether measurement errors only occur with age or also with duration of residence, but there is a clear difference between the two cases with measurement error and the estimate based on the original data. While the sum of the absolute values of the deviation between the estimator and the true value is around 285 across all areas for the true data, this value is almost 400 in the two cases with measurement error. This represents an increase of more than one third and shows why it would make sense to use the measurement error Fay-Herriot model from the previous chapter. For the second case, where the duration of residence is also subject to measurement error, the results of the estimation are also looked at in the Horvitz-Thompson estimator and the post-stratified estimator. The results of this are shown in table 7.9.

Estimator $\hat{\bar{Y}}$	**Sum of deviations**	**Sum of absolute deviations**	**Variance per area**	**Variance per area in the small areas**
Horvitz-Thompson estimator **Original** data	- 49.16	166.93	8.80	52.72
Horvitz-Thompson estimator **Measurement error**	- 37.38	432.13	8.17	47.48
Post-stratified estimator **Original** data	- 172.24	256.11	1.16	4.20
Post-stratified estimator **Measurement error**	- 145.83	399.54	1.14	3.94

Table 7.9: Deviations from the true mean and variances of HT estimator and PSE for data with measurement error

Similar to the MSE for the model-based estimators, only minimal differences are seen here in the variance of the estimators between the original data and the data with measurement errors. In both cases, the average variance per city district quarter decreases marginally in all areas as well as in the small areas. On the other hand, as in the case of the Fay-Herriot estimator, the deviations of the estimator from the true value also show larger differences in these two estimators. The Horvitz-Thompson estimator shows the largest increase. The deviations over the summed absolute values of all areas amount to more than 430 for the data with measurement error and are thus far more than double those of the estimate with the original data. This shows that the Horvitz-Thompson estimator is the most susceptible of the four estimators examined here with respect to measurement error. With the post-stratified estimator, this sum increases from around 250 to 400, similar to the Fay-Herriot estimator. It can thus be seen that the Horvitz-Thompson estimator on the original data has the least bias of these four estimators, whereas it yields the highest value of summed deviations in the case of measurement error.

This finding concludes the practical analysis of the different small area estimation methods on the Munich population data and deficit versions thereof. Interesting and sometimes surprising results were obtained in many places. The most important of these will be discussed again in the next concluding chapter.

Chapter 8

Conclusion and outlook

This chapter gives a final conclusion about this thesis and takes up the most important results of all chapters before giving a short outlook.

Conclusion

Whether it is purchase prices for building land, poverty measurement, data on Covid or business statistics - in more and more areas of the economy, politics and society, data is analysed at small spatial levels in order to get the most accurate overview of the situation. The basis for this is small area estimation, with whose methods the described facts can be analysed on a small scale. The diverse current fields of application of small area estimation show the relevance of small area estimation in modern statistics and this is practically the foundation of this thesis and this conclusion is also based on that.

However, in order to be able to apply spatial small area methods at all, georeferenced data is required, which is provided with a spatial reference and can be assigned to a location on earth. Therefore, georeferencing is also an important part of this thesis. Georeferencing was considered as the first step of spatial data analysis, where georeferenced data is transformed into a desired reference system in the context of geocoding. One main difference here is between formal and informal georeferencing. In the former, exact spatial coordinates are assigned to the data and in the latter, the assignment is done via colloquial references such as street names. A distinction is also made between vector and raster referencing. With vector referencing, the locations are assigned to an element in a digital map via a vector such as points or lines. With raster referencing, pixels in a raster image are provided with a geographical reference. The spatial references are also well suited for linking various data sets from different topics. So-called grids, which play an important role especially in georeferencing in official statistics, are also suitable for this purpose. Due to a renewal of the Federal Statistics Act in August 2013, it has since been possible to permanently store the spatial reference in a grid with a minimum width of 100 metres before deleting the exact address of the respondent. This ensures a good protection of the data and at the same time gives the possibility for spatial analyses. Examples of the use of georeferenced data in official statistics are the census atlas or accessibility analyses such as the hospital atlas.

The presence of spatial references allows the total population of a dataset to be divided into several subpopulations in which the estimation of certain parameters and key figures is of interest. Due to the spatial subdivision, in many areas the sample size is so small that direct estimators like the Horvitz-Thompson estimator lead to imprecise estimates with too high variance. Therefore, various estimation methods have been introduced in this thesis that correspond to the core of small area estimation and help to provide precise estimates even for small-scale areas. These are mainly indirect estimators which, under the principle of borrowing strength, also take into account information from other areas in addition to auxiliary variables. Since the estimators presented in this thesis were used in practice to estimate the duration of residence on the basis of population data from the Munich statistical office, the results of the estimators are taken into account in the duration of residence estimation in this conclusion. Both the design-based estimators with the basic synthetic estimator and the post-stratified estimator as well as the sample size dependent estimator as a combination of the two and the generalised regression estimator and the area-specific generalised regression estimator as model-assisted estimators show a considerably smaller variance in the estimation of the duration of residence, especially in areas with a small sample size, than the Horvitz-Thompson estimator. Since the individual age-sex groups differ too much with regard to the mean duration of residence, the post-stratified estimator turns out to be a much better estimator than the basic synthetic estimator in the present case. The area-specific generalised regression estimator shows more absolute deviations of the estimates from the true value than the generalised regression estimator, but has a considerably smaller variance.

In addition to these estimators, the Battese-Harter-Fuller estimator as a unit-level model and the Fay-Herriot estimator at area-level were also applied as model-based estimators to the population data as the central estimators of small area estimation. They are based on mixed models in order to take into account the dependence of the elements within an area through random effects. Furthermore, they fulfil the property that they are the (empirical) best linear unbiased predictors. Moreover, their great importance within small area estimation has been demonstrated by their numerous uses in application examples of official statistics. Since the population data are individual data, the Battese-Harter-Fuller estimator can be applied, as well as the Fay-Herriot estimator when aggregating the data. However, it cannot be determined exactly which estimator is better suited for estimating the durations of residence. The sum of the deviations from the true value does not differ much between the two estimators and the average mean squared error for all city district quarters is smaller for the Battese-Harter-Fuller estimator and in the city district quarters with a sample size below 50 the Fay-Herriot estimator has a smaller MSE. In addition, there are numerous spatial, temporal, multivariate and generalised extensions of these two estimators.

A major focus of this work is also on the deficient data. There are not only missing data but also many other forms of coarse data such as outliers, measurement errors or rounding. With regard to official statistics, confidentiality methods are also relevant, which comply with the principle of safekeeping of data and therefore modify the data before publication. Thus, small values of so-called risky cells can be completely suppressed or the data can be changed by noise addition or data swapping. To deal with these types of deficient data, a robust small area estimation approach was looked at to better handle outliers. Cautious modelling without strong incredible assumptions is used to deal with missing data and a bivariate Fay-Herriot model was presented that is specific to measurement error. In the practical analysis when looking at residential duration estimation for different types of deficient data, it became clear once again why direct estimators like the Horvitz-Thompson estimator are problematic in small area estimation. For example, when missing data occurred, its variance was again much larger than with the original data. In contrast, the Battese-Harter-Fuller estimator, for example, has shown very good properties in the presence of measurement errors, which is why the use of the two model-based estimators has been worthwhile at many points in this thesis. Moreover, in the case of missing and imputed data, a very similar behaviour of the small area estimation was found for the population data compared to a simulated dataset.

Outlook

The many examples of applications of georeferencing and small area estimation in official statistics, some of which have been presented in several parts of this thesis, have already shown the actuality of this topic. However, these issues will continue to be present in the future. During the EMOS Day on 25 November 2022 at the Bavarian Statistical Office in Fürth, projects were presented in many different subject areas on how small area estimation will be used even more in official statistics in the future. In the context of Big Data, even more data with georeferences will be available for this purpose in the future, as advancing technology will make even more possibilities for data generation accessible. Although this paper has already presented some extensions of the model-based estimators that take into account the positive properties of the Battese-Harter-Fuller estimator and the Fay-Herriots estimator, there is still much room for research here to develop even more new estimators that adapt to the conditions of technical progress. Especially when it comes to the occurrence of deficient data in the context of small area estimation, there is still a lot of room in this regard and further modelling approaches would be helpful here. Particularly in the context of Big Data, data quality is often deficient, which must not be neglected in future in small area estimation with data from Big Data sources.

Bibliography

- Blazquez, Desamparados and Josep Domenech (2018): Big Data sources and methods for social and economic analyses, Technological Forecasting and Social Change, Vol. 130, pp. 99-113, `https://doi.org/10.1016/j.techfore.2017.07.027` (last visit: 22/02/2023), Valencia.

- Bohnensteffen, Sarah, Mühlhan, Jannek and Younes Saidani (2021): Mobilität während der Corona-Pandemie, in Statistisches Bundesamt (Editor): WISTA – Wirtschaft und Statistik, 3/2021, pp. 89-105, `https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Stati` `2021/03/mobilitaet-corona-pandemie-032021.pdf?__blob=publicationFile` (last visit: 22/02/2023), Wiesbaden.

- Brenner, Carmina (2015): Raumbezug in der amtlichen Statistik und Anwendungsbeispiele, in: Statistisches Monatsheft Baden-Württemberg, 9/2015, pp. 47-53, `https://www.statistik-bw.de/Service/Veroeff/Monatshefte/PDF/Beitrag15_09_08.pdf` (last visit: 22/02/2023), Stuttgart.

- Brenzel, Hanna and Kathrin Gebers (2020): Werkstattbericht: Georeferenzierung im statistischen Verbund, in Statistisches Bundesamt (Editor): WISTA – Wirtschaft und Statistik, 6/2020, pp. 48-57, `https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statis` `2020/06/werkstattbericht-062020.pdf?__blob=publicationFile` (last visit: 22/02/2023), Wiesbaden.

- Bundesamt für Kartographie und Geodäsie (2023): Zeitplan – Meilensteine bei der Umsetzung der INSPIRE-Richtlinie, `https://www.gdi-de.org/INSPIRE/Zeitplan` (last visit: 22/02/2023), Frankfurt am Main.

- Bundesamt für Landestopografie swisstopo (2023): Geoinformation und Geodaten, `https://www.swisstopo.admin.ch/de/wissen-fakten/geoinformation.html` (last visit: 22/02/2023), Wabern.

- Bundesverband Geothermie e.V. (2020): Georeferenzierte Daten, Lexikon der Geothermie, `https://www.geothermie.de/bibliothek/lexikon-der-geothermie/g/georeferenzierte-date` `html` (last visit: 22/02/2023), Berlin.

- Burgard, Jan Pablo, Esteban, Maria Dolores, Morales, Domingo and Agustin Perez (2021): Small area estimation under a measurement error bivariate Fay–Herriot model, Statistical Methods & Applications (30), pp. 79-108, `https://doi.org/10.1007/s10260-020-00515-9` (last visit: 22/02/2023).

- Chambers, Raymond L. (1986): Outlier Robust Finite Population Estimation, Journal of the American Statistical Association, Vol. 81, No. 396 (December 1986), pp. 1063-1069, `https://www.jstor.org/stable/2289084` (last visit: 22/02/2023).

- Duncan, George T., Elliot, Mark and Juan-José Salazar-González (2011): Statistical Confidentiality, Principles and Practice, 1st edition, Springer, https://doi.org/10.1007/978-1-4419-7802-8 (last visit: 22/02/2023), New York.

- Eurostat (2018): Verhaltenskodex für europäische Statistiken, Europäische Union, revised edition 2017, https://ec.europa.eu/eurostat/documents/4031688/9394019/KS-02-18-142-DE-N.pdf/27ca19ca-e349-45f8-bbd4-4d78a33601ae?t=1542709797000 (last visit: 22/02/2023), Luxemburg.

- Fahrmeir, Ludwig, Lang, Stefan, Kneib, Thomas and Brian Marx (2013): Regression, Models, Methods and Applications, Springer-Verlag Berlin Heidelberg, DOI: 10.1007/978-3-642-34333-9, https://link.springer.com/book/10.1007/978-3-642-34333-9 (last visit: 22/02/2023), Heidelberg.

- Faulbaum, Frank (2014): Total Survey Error, in Baur, Nina and Jörg Blasius (Editors): Handbuch Methoden der empirischen Sozialforschung, pp. 439-453, Springer Fachmedien, https://link.springer.com/book/10.1007/978-3-531-18939-0 (last visit: 22/02/2023), Wiesbaden.

- Fay, Robert E. and Roger A. Herriot (1979): Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data, Journal of the American Statistical Association, Jun. 1979, Vol. 74, No. 366, pp 269-277, https://www.jstor.org/stable/2286322 (last visit: 22/02/2023).

- Gebers, Kathrin and Philip Graze (2019): Statistische Datengewinnung durch die Nutzung geografischer Informationen, in Statistisches Bundesamt (Editor): WISTA – Wirtschaft und Statistik, 4/2019, pp. 11-18, https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2019/04/statistische-datengewinnung-042019.pdf?__blob=publicationFile (last visit: 22/02/2023), Wiesbaden.

- GPS Koordinaten (2023): GPS Koordinaten – Breiten- und Längengrad einer Adresse, https://www.gpskoordinaten.de/ (last visit: 22/02/2023).

- Groves, Robert M. and Lars Lyberg (2010): Total Survey Error: Past, Present, and Future, Public Opinion Quarterly, Volume 74, Issue 5, pp. 849–879, https://doi.org/10.1093/poq/nfq065 (last visit: 22/02/2023), Oxford.

- Hackeloeer, Andreas, Klasing, Klaas, Krisp, Jukka M. and Liqiu Meng (2014): Georeferencing: a review of methods and applications, Annals of GIS, 20:1, pp. 61-69, DOI: 10.1080/19475683.2013.868826, https://doi.org/10.1080/19475683.2013.868826 (last visit: 22/02/2023).

- Hastings, Jordan T. and Linda L. Hill (2018): Georeferencing, in: Liu, L. and M.T. Özsu (Editor), Encyclopedia of Database Systems, pp. 1616-1620, Springer, https://doi-org.emedien.ub.uni-muenchen.de/10.1007/978-1-4614-8265-9_181 (last visit: 22/02/2023), New York.

- Heitjan, Daniel F. and Donald B. Rubin (1991): Ignorability and Coarse Data, The Annals of Statistics, Vol. 19, No. 4, pp. 2244-2253, https://www.jstor.org/stable/2241929 (last visit: 22/02/2023), Ann Arbor, Michigan.

- Hobza, Tomas, Morales, Domingo, Esteban, Maria Dolores and Augustin Perez (2021): A Course on Small Area Estimation and Mixed Models, Methods, Theory and Applications in R, 1st Edition, Springer Nature Switzerland AG, https://doi.org/10.1007/978-3-030-63757-6 (last visit: 22/02/2023), Cham.

- Islam, M. Ataharul and Rafiqul I. Chowdhury (2017). Analysis of Repeated Measures Data, Springer Nature Singapore Pte Ltd., DOI: 10.1007/978-981-10-3794-8, https://link.springer.com/book/10.1007/978-981-10-3794-8 (last visit: 22/02/2023), Singapore.

- Kleinke, Kristian, Reinecke, Jost, Salfrán, Daniel and Martin Spiess (2020): Applied Multiple Imputation, Advantages, Pitfalls, New Developments and Applications in R, Statistics for Social and Behavioral Sciences, Springer Nature Switzerland AG, https://doi.org/10.1007/978-3-030-38164-6 (last visit: 22/02/2023), Cham.

- Landesbetrieb Information und Technik Nordrhein-Westfalen (2023): Small Area-Methode zur Schätzung von durchschnittlichen Bestandsmieten auf Gemeindeebene, Experimentelle Statistik – Neue Datenquelle/Methode, Statistik und IT-Dienstleistungen, https://www.it.nrw/small-area-methode-zur-schaetzung-von-durchschnittlichen-bestandsmi
(last visit: 22/02/2023), Düsseldorf.

- Lehtonen, Risto and Ari Veijanen (2016): Estimation of Poverty Rate and Quintile Share Ratio for Domains and Small Areas, in Alleva, Giorgio and Andrea Giommi (Editors): Topics in Theoretical and Applied Statistics, pp. 153-165, Springer International Publishing Switzerland, DOI: 10.1007/978-3-319-27274-0, https://link.springer.com/book/10.1007/978-3-319-27274-0 (last visit: 22/02/2023), Cham.

- Little, Roderick J. A. and Donald B. Rubin (2002): Statistical Analysis with Missing Data, Second Edition, Wiley Series in Probability and Statistics, ISBN 0-471-18386-5, John Wiley & Sons, Hoboken, New Jersey.

- Manecke, Julia (2019): Regionale Auswertung von Unternehmensstatistiken: Methoden und Anwendungen im Kontext der Small-Area-Statistik, in Statistisches Bundesamt (Editor): WISTA – Wirtschaft und Statistik, 1/2019, pp. 143-152, https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2019/01/auswertung-unternehmenss
pdf?__blob=publicationFile (last visit: 22/02/2023), Wiesbaden.

- Molina, Isabel and Yolanda Marhuenda (2015): sae: An R Package for Small Area Estimation, The R Journal, Vol. 7/1, June 2015, pp. 81-98, https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf (last visit: 22/02/2023).

- Münnich, Ralf, Burgard, Jan Pablo and Martin Vogt (2013): Small Area-Statistik: Methoden und Anwendungen, in: Deutsche Statistische Gesellschaft (2013): AStA Wirtsch Sozialstat Arch, 6:149-191, DOI: 10.1007/s11943-013-0126-1, https://doi.org/10.1007/s11943-013-0126-1 (last visit: 22/02/2023), Trier.

- Münnich, Ralf, Burgard, Jan Pablo, Ertz, Florian, Lenau, Simon, Manecke, Julia and Hariolf Merkle (2019): Guidelines on small area estimation for city statistics and other functional geographies, 2019 Edition, European Union, Publications Office of the European Union, DOI: 10.2785/822325, https://ec.europa.eu/eurostat/de/web/products-manuals-and-guidelines/-/ks-gq-19-011(last visit: 22/02/2023), Luxembourg.

- Neutze, Michael (2015): Gitterbasierte Auswertungen des Zensus 2011, in: Verband Deutscher Städtestatistiker (Editor), Stadtforschung und Statistik, 2/2015, pp. 64-67, https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Aufsaetze_Archiv/2015_02_Destatis_GitterbasierteAuswertungen.pdf?__blob=publicationFile&v=2 (last visit: 22/02/2023), Stuttgart.

- Plass, Julia, Omar, Aziz and Thomas Augustin (2017): Towards a Cautious Modelling of Missing Data in Small Area Estimation, PMLR: Proceedings of Machine Learning Research, Vol. 62, 2017, pp. 253-264, https://proceedings.mlr.press/v62/plass17a.html (last visit: 22/02/2023).

- Polettini, Silvia and Serena Arima (2015): Small area estimation with covariates perturbed for disclosure limitation, Statistica, 75 (1), pp. 57-72, https://doi.org/10.6092/issn.1973-2201/5823 (last visit: 22/02/2023), Rome.

- Professur für Geodäsie und Geoinformatik (2001): Geokodierung, Geoinformatik-Service, Universität Rostock, http://www.geoinformatik.uni-rostock.de/einzel.asp?ID=760 (last visit: 22/02/2023), Rostock.

- Radermacher, Walter J. (2019): Official Statistics 4.0, Verified Facts for People in the 21st Century, 1st edition, https://doi.org/10.1007/978-3-030-31492-7 (last visit: 22/02/2023), Springer Nature Switzerland AG, Cham.

- Rao, J.N.K. and Isabel Molina (2015), Small Area Estimation, Second Edition, JohnWiley&Sons, Inc., DOI: 10.1002/9781118735855, https://doi.org/10.1002/9781118735855 (last visit: 22/02/2023), Hoboken, New Jersey.

- Ribeiro, Ana Isabel, Olhero, Andreia, Teixeira, Hugo, Magalhaes, Alexandre and Maria Fatima Pina (2014): Tools for Address Georeferencing – Limitations and Opportunities Every Public Health Professional Should Be Aware Of, PLoS ONE 9(12): e114130, https://doi.org/10.1371/journal.pone.0114130 (last visit: 22/02/2023), San Francisco.

- Särndal Carl-Erik and Jean-Claude Deville (1992): Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, Vol. 87, No. 418 (Jun 1992), pp. 376-382, https://www.jstor.org/stable/2290268 (last visit: 22/02/2023).

- Schnorr-Bäcker, Susanne (2012): Georeferenzierung von Daten, Zum gleichnamigen Abschlussbericht der Arbeitsgruppe „Georeferenzierung von Daten“ des Rates für Sozial- und Wirtschaftsdaten aus Sicht der Bundesstatistik, in Statistisches Bundesamt (Editor): WISTA – Wirtschaft und Statistik, July 2012, pp. 563-571, https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2012/07/georeferenzierung-dat pdf?__blob=publicationFile (last visit: 22/02/2023), Wiesbaden.

- Sinha, Sanjoy K. and J.N.K. Rao (2009): Robust small area estimation, The Canadian Journal of Statistics, Vol. 37, No. 3 (September 2009), pp. 381-399, https://www.jstor.org/stable/25653486 (last visit: 22/02/2023).

- Spektrum der Wissenschaft Verlagsgesellschaft mbH (2023): Georeferenzierung, Lexikon der Geowissenschaften, Spektrum Akademischer Verlag, https://www.spektrum.de/lexikon/geowissenschaften/georeferenzierung/5724 (last visit: 22/02/2023), Heidelberg.

- Statistische Ämter des Bundes und der Länder (2022): Regionaldatenbank Deutschland, Kaufwerte für Bauland, Tabelle 61511-01-03-4, https://www.regionalstatistik.de/genesis//online?operation=table&code=61511-01-03-4&bypass=true&levelindex=1&levelid=1655808031723#abreadcrumb (last visit: 22/02/2023), Düsseldorf.

- Statistisches Bundesamt (2019): Digitale Agenda des Statistischen Bundesamtes, 03/2019, Version 2.1, https://www.destatis.de/DE/Service/OpenData/Publikationen/digitale-agenda.pdf?__blob=publicationFile (last visit: 22/02/2023), Wiesbaden.

- United States Census Bureau (2022): About SAIPE, https://www.census.gov/programs-surveys/saipe/about.html (last visit: 22/02/2023), Washington DC.

- Ushakov, N.G. and V.G. Ushakov (2018): Statistical Analysis of Rounded Data: Measurement Errors vs Rounding Errors, Journal of Mathematical Sciences volume 234, pp. 770–773, https://doi.org/10.1007/s10958-018-4042-3 (last visited: 22/02/2023).

- Zhao, Ningning and Zhidong Bai (2012): Analysis of rounded data in mixture normal model, Stat Papers 53, pp. 895–914, https://doi.org/10.1007/s00362-011-0395-0 (last visit: 22/02/2023).

R Packages

The following R packages are used for the practical part of this thesis:

mice van Buuren, Stef and Karin Groothuis-Oudshoorn (2011): mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, 45(3), 1-67, DOI: 10.18637/jss.v045.i03, `https://www.jstatsoft.org/article/view/v045i03` (last visit: 22/02/2023).

rgdal Bivand R, Keitt T and B Rowlingson (2022): “rgdal: Bindings for the ’Geospatial’ Data Abstraction Library“, R package version 1.6-3, `https://CRAN.R-project.org/package=rgdal` (last visit: 22/02/2023).

sae Molina I and Y Marhuenda (2015): “sae: An R Package for Small Area Estimation.“, The R Journal, *7*(1), 81-98. `https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf` (last visit: 22/02/2023).

sf Pebesma, E (2018): Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, `https://doi.org/10.32614/RJ-2018-009` (last visit: 22/02/2023).

sp Pebesma, E.J. and R.S. Bivand (2005): Classes and methods for spatial data in R. R News 5 (2), `https://cran.r-project.org/doc/Rnews/` (last visit: 22/02/2023).

tidyverse Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Mueller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K and H Yutani (2019): “Welcome to the tidyverse.“, Journal of Open Source Software, *4*(43), 1686, doi:10.21105/joss.01686, `https://doi.org/10.21105/joss.01686` (last visit: 22/02/2023).

tmap Tennekes, M (2018): “tmap: Thematic Maps in R.“, Journal of Statistical Software, *84*(6), 1-39. doi:10.18637/jss.v084.i06, `https://doi.org/10.18637/jss.v084.i06`, (last visit: 22/02/2023).

tmaptools Tennekes, M (2021): “tmaptools: Thematic Map Tools“, R package version 3.1-1, `https://CRAN.R-project.org/package=tmaptools` (last visit: 22/02/2023).

Digital appendix

The attached digital storage medium has the following content:

- This master thesis in PDF format
- The folder “Graphics“ with all graphics of this thesis
- The R script with the code for the practical analysis of this thesis

The data set used with the Munich population data may not be passed on for data protection reasons and is therefore not on the storage medium.

List of Figures

List of Tables

Acknowledgement

At the end of this thesis I would like to thank Professor Dr Thomas Augustin not only for the supervision during this thesis but also during the whole EMOS master. He has always been an important, trustworthy and helpful contact person for questions regarding the EMOS programme, the associated lectures and the internship. Thank you very much!

Furthermore, I would like to thank all the people at the Institute of Statistics at LMU Munich who took care of the EMOS Master, especially Dominik Kreiss. I would also like to thank the members of the Statistical Office of the City of Munich for supervising my EMOS internship and for providing the population data for this thesis.

The following part of this acknowledgement should please not be taken too seriously, but is nevertheless close to my heart:
Ois Oablinger Bua woid i doch no a paar Worte loswern, de ned auf Englisch han, um mi bei dejenigen zum bedanken, de mi außerhoib von da Uni während da gesamten Zeit unterstützt ham. Ois Ersts guit mei Dank meine Eltern und meiner Familie. Meiner Mama für de mentale Unterstützung und dafür, dass i oiwei guad versorgt war während der Zeit, weil da Geist nur mit guadm Essen arban ko. Meim Vater fürs Durchlesn vo jegliche Arbeitn während am Studium. Und meiner Oma, de uns leider während da Masterarbeit verlassn hod und nur no vo om zuaschaung ko, aber trotzdem oiwei für mi do gwen is.

Außerdem woid i mi bei jedm vo meine Freind bedanken. Speziell a bei da Laufgruppe in Oabling, de immer für Abwechslung während da Masterarbeit gsorgt hod und mit der i während des gesamten Masters so vui scheene Sachan während am Training, de Wettkämpfe oder in die Berg erlem hob derfen. Des gleiche guid a fürs Happy Fitness Laufteam aus Innsbruck und am PTSV Rosenheim, speziell für die Endzeit vom Master. A Dank guid a Maxlroa und da Essbar für de Arbeisstelln während de Semesterferien und da Studiumszeit.

Dank eich für ois!!!

Declaration of Originality

I confirm that the submitted master thesis with the title

Small area estimation and georeferencing in the context of deficient data - Methodology and applications with data from the Munich municipal statistics

is original work and was written by me without further assistance. Appropriate credit has been given where reference has been made to the work of others. The master thesis was not examined before, nor has it been published.

Anian Rottmüller

Bad Aibling, February 22, 2023