Counting the Uncounted

Methodological Extensions in Multiple Systems Estimation

Enkele Uitbreidingen in Multiple Systems Estimation
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 22 november 2024 des ochtends te 10.15 uur

door

Daan Bernardus Zult

geboren op 7 oktober 1979

te Hoorn

Promotoren:
Prof. dr. P.G.M. van der Heijden
Prof. dr. B.F.M. Bakker

Beoordelingscommissie:
Prof. dr. K. van Deun
Dr. P.J. Lugtig
Prof. dr. D.L. Oberski
Prof. dr. B. Schouten
Prof. P.A. Smith

To Sylvia, Olav and Nova, I am so lucky to have you in my life!

Zeg, zal ik je eens even wat vertellen

Zeg, zal ik je eens even wat vertellen
De grote mensen – ook al is dat raar
Die denken dat ze alles kunnen tellen
Maar weet je – dat is helemaal niet waar

Vraag ze hoeveel sprietjes je kunt vinden in het gras
Vraag ze hoeveel druppels je kunt vangen in een glas
Vraag ze hoeveel sterren boven aan de hemel staan
En hoeveel mensen snurken als ze 's avonds slapen gaan.

Zeg, zal ik je eens even wat vertellen
De grote mensen – ook al is dat raar
Die denken dat ze alles kunnen tellen
Maar weet je – dat is helemaal niet waar

Ieniemienie - *Sesamstraat, Vriendjes voor altijd* (2007).

The writing of this dissertation was supported by Statistics Netherlands, but responsibility of the views expressed as well as errors or omissions made, belong solely to the author(s), and do not reflect the views of Statistics Netherlands.

Contents

Chapter 1

Introduction

How to count the uncounted? This question can be interpreted both literally and figuratively. First, how to count those that literally cannot be counted, simply because they are unobserved? Second, how to count those who figuratively do not seem to count, such as marginalised or hard-to-reach groups? Multiple Systems Estimation (MSE), a statistical model that is designed to deal with both cases, is the main topic of this dissertation.

The traditional MSE setting is the literal case, where a population of interest is not fully observed by one complete list that contains one record for each population unit, but by two or more samples that each contain a different subset of this population. Then some population units may be present in one or more samples and some may not be present in any of the samples at all. MSE is designed to provide an estimate for this missing part. This traditional setting mainly stems from the field of ecology, where the estimation of the size of animal populations plays a major role (see e.g. Petersen, 1896; Lincoln, 1930). This setting may also arise in human populations when individuals are not accurately administered, therefore MSE is nowadays also applied by National Statistical Institutes (NSIs) to estimate the so-called undercoverage of their census (see e.g. Hogan, Cantwell, Devine, Mule, & Velkoff, 2013; Wolter, 1986) or person register (see e.g. Bakker, van Rooijen, & van Toor, 2014; Statistic Netherlands, 2016; Bakker, van der Heijden, & Gerritse, 2017).

The second case of marginalised or hard-to-reach groups starts from a broader interpretation of being uncounted, because it also concerns the case where a complete list or register of the population may be available, but some information of interest is missing. An illustrative example of such a case that plays an important role in this dissertation, is the question of how many people in The Netherlands are homeless (see also Coumans, Cruyff, van der Heijden, Wolf, & Schmeets, 2017)? Similar cases of the use of MSE to estimate the size of hard-to-reach groups can be found in epidemiology, where MSE is used to estimate how many people or animals carry some sort of disease (see e.g. Gill, Ismail, Beeching, Macfarlane, & Bellis, 2003; Muneza et al., 2017; Böhning, Rocchetti, Maruotti, & Holling, 2020) or use some sort of drug (see e.g White, Bird, & Grieve, 2014). Also in the domain of public policy and human rights research (see e.g. Lum, Price, & Banks, 2013), MSE is used to estimate the number of war casualties (see e.g. Manrique-Vallier, Price, & Gohdes, 2013), the number of people that are a victim of human trafficking (see e.g. UNODC, 2022), forced labour (see e.g. Belser, de Cock, Mehran, & ILO, 2005; ILO, 2018) or modern slavery (see e.g. Silverman, 2020; Binette & Steorts, 2022). In van der Heijden et al. (2021) it is used

to estimate the size of the Māori population in New Zealand, in Yauck, Rivest, and Rothman (2019) it is used to estimate the number of visitors to a business location and Fienberg, Johnson, and Junker (2002) even use it to estimate the size of the World Wide Web. A wide overview of some of these and other applications are discussed in International Working Group for Disease Monitoring and Forecasting (1995b); Bird and King (2018).

As Chao (2015) writes, the basic idea of MSE with two samples can be traced back to a 1786 paper by Pierre Simon Laplace, who used it to estimate the population size of France in 1802 (Cochran, 1978; Seber, 1982), and even earlier to John Graunt who used the idea to estimate the effect of plague on the population size of England around 1600 (Hald, 1975). The theoretical development that led to modern MSE theory, took off in the field of ecology with the work of Petersen (1896); Lincoln (1930) and Schnabel (1938). It became more generally known and more widely applicable due to the work by Sekar and Deming (1949), who used it to estimate the size of a human population. Jolly (1965) and Seber (1965) proposed their Jolly-Seber model for populations that could be subject to events such as death, birth and migration. Later, Bishop, Fienberg, and Holland (1975) strengthened its theoretical foundation further by establishing a link between MSE and the log-linear model. A discussion of the relation between the model and underlying assumptions was provided by e.g. Wolter (1986), who extensively discussed the relation between the two sample estimator and its underlying assumptions. A more comprehensive description of the history of MSE and its theoretical development can be found in a paper by the International Working Group for Disease Monitoring and Forecasting (1995a).

Over time, MSE became known under different names that all refer to the same method. The name that is used generally may depend on the number of samples that is involved and on the scientific field in which it is discussed. When two samples are involved, in ecology the method is usually referred to as Capture-recapture (see e.g. Amstrup, McDonald, & Manly, 2005) or Mark-recapture (see e.g. McClintock, Conn, Alonso, & Crooks, 2013). In other fields it is also referred to as Dual-system estimation (see e.g. Cantwell, 2014). When there are $k > 2$ samples, also the name Multiple-recapture (see e.g. Darroch, 1958; Cormack, 1989) estimation is encountered and when there are exactly three samples involved the method may be called Triple-system estimation (see e.g. Zaslavsky & Wolfgang, 1993; Baffour, Brown, & Smith, 2013). In the different chapters of this dissertation these different names are used somewhat interchangeably.

Despite its long history in scientific literature, MSE still faces some unresolved theoretical and practical problems. These issues also play a role in the practice of producing population size statistics at Statistics Netherlands, which is the driving force behind the work presented in this dissertation. These issues are of a general nature and may therefore be interesting to any MSE practitioner. In the remainder of this introduction we discuss some of these unresolved theoretical and practical MSE issues, and explain how they are related to the different chapters. Before we can discuss these relations, we first need to introduce some of the methodological basics

of MSE. Finally, this chapter summarises the conclusions of the different chapters, discusses open issues and suggests some further research.

1.1 The basics of multiple systems Estimation

The basic idea of MSE is that the size of a population can be estimated by combining different samples from this population. This idea can be most clearly illustrated by discussing in some further detail the simple case of a population that is partly observed by two samples that we will refer to as dual-system estimation (DSE). DSE assumes a population with size N and two samples A and B that each contain a random sample from this population. It is assumed that each population unit can be perfectly identified and therefore it is possible to count the number of unique population units in sample A, the number of unique population units in sample B and the number of unique population units in both samples. These counts can be denoted as n_{ab} with $a \in (1, 0, +)$ where $a = 1$ means *in sample A*, $a = 0$ means *not in sample A* and $a = +$ means *both in and not in sample A*, and the same for b. This notation gives n_{1+} as the size of sample A, n_{+1} as the size of sample B and n_{00} as the unobserved part of the population. These counts can be presented more schematically as in Table 1.1.

Table 1.1: Illustration of the problem with two samples

Sample A \ Sample B	in B	not in B	both in and not in B
in A	n_{11}	n_{10}	n_{1+}
not in A	n_{01}	n_{00} =?	?
both in and not in A	n_{+1}	?	N

Table 1.1 shows that if n_{00} is known, then $N = n_{11} + n_{10} + n_{01} + n_{00}$ would be known as well. Of course, the problem is that n_{00} is not observed. Table 1.1 suggests some intuition of how n_{00} may be estimated as well. When it is assumed that the probability of a population unit being included in sample A is independent of this unit being included in sample B, Table 1.1 suggests that the ratio n_{11}/n_{01} in the first column should, on average, be approximately equal to the ratio n_{10}/n_{00} in the second column. When we consider n_{00} to be a random variable with expectation m_{00}, this reasoning directly suggests a DSE-estimator for m_{00} that can be written as

$$\hat{m}_{00}^{\text{LP}} = n_{10} n_{01} / n_{11}. \tag{1.1}$$

This DSE-estimator was already proposed by Petersen (1896), and later Lincoln (1930), and is therefore also often referred to as the Lincoln-Petersen (LP) estimator. Later, Chapman (1951) introduced an alternative but very similar DSE-estimator, i.e.:

$$\hat{m}_{00}^{\text{Chapman}} = n_{10} n_{01} / (n_{11} + 1), \tag{1.2}$$

which has better small sample properties.

An important role in much of the work presented in this dissertation is related to the work by Fienberg (1972), who showed that MSE-estimators and therefore also DSE-estimators can be obtained from a log-linear model. For the LP-estimator this starts with the log-linear model equation

$$\log m_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB}, \tag{1.3}$$

with m_{ab} the expectation of n_{ab}, λ an intercept term, λ_a^A and λ_b^B are the respective inclusion parameters for sample A and B that are identified by setting $\lambda_0^A = \lambda_0^B = 0$ and λ_{ab}^{AB} is a parameter for the interaction between sample A and B. For m_{00} this reduces to $m_{00} = \exp\lambda$. Because m_{00} is unobserved it is usually assumed that $\lambda_{ab}^{AB} = 0$. Then, Eq. (1.3) reduces to three equations and three unknowns that, when solved for λ, give the LP-estimator in Eq. (1.1).

The assumptions under which $\hat{m}_{00}^{\mathrm{LP}}$ is an asymptotically unbiased estimator are discussed by Wolter (1986); International Working Group for Disease Monitoring and Forecasting (1995a); Zhang (2019) and others. We will briefly discuss them here as well, because they lead to the topics dealt with in the chapters of this dissertation. The four assumptions can be outlined as follows:

1. The sampling population is equal for both samples.
2. In each sample, population units can be perfectly identified.
3. Inclusion probabilities are homogeneous in at least one of the samples.
4. The samples are independent.

When each of these assumptions hold and the samples are of sufficient size (i.e. when $n_{1+}n_{+1}/N > \log N$, (Chapman, 1951), which could be seen as a fifth assumption or a regularity condition), one can apply DSE without too much concern. When one of these assumptions is violated and not taken into account appropriately, both the LP- and Chapman-estimator will be biased.

Whether and which of these assumptions is violated depends on the application. When DSE is applied on data available to a NSI, it is very common that both DSE assumption 3 and 4 are violated. Assumption 3 is often unlikely to hold, because in public administration inclusion probabilities generally differ between different groups. For example, older people have a larger probability to be included in a hospital register, while younger people have a larger probability to be included in an education register. A standard way to deal with these different probabilities is to add categorical covariates to the model. For example, one may construct Table 1.1 for young and old people separately and estimate an m_{00} for each group. Another way that leads to the same result is to add categorical covariate parameters to the log-linear model.

Assumption 4 is also often violated, simply because a person being included in one register may affect the probability that this person is included in another register.

This might be due to some unobserved personal factor that increases the probability of a person being (or not being) registered in any register, but also due to administrations that use the other register to improve their own coverage. With animal populations, the correlation between samples is often explained by animals that after being trapped, become either more "trap happy" or more "trap shy". In the context of Eq. (1.3), this implies that the assumption $\lambda_{ab}^{AB} = 0$ is violated and so $m_{00} = \exp \lambda$ no longer holds. A standard solution to this violation of the independence assumption 4, is to use more than two samples (Fienberg, 1972). For example, when three sample A, B and C are available, the number of observed counts is seven ($n_{111}, n_{110}, n_{101}, n_{011}, n_{100}, n_{010}$ and n_{001}), which allows the log-linear model to be extended to:

$$\log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ab}^{AB} + \mu_{ac}^{AC} + \mu_{bc}^{BC}. \tag{1.4}$$

Eq. (1.4) consists of seven equations and seven parameters. It does not contain μ_{abc}^{ABC}, because for identification it is assumed that $\mu_{abc}^{ABC} = 0$. The main feature of Eq. (1.4) as compared to Eq. (1.3), is that the interaction parameters μ_{ab}^{AB}, μ_{ac}^{AC} and μ_{bc}^{BC} can also be estimated, hereby controlling for pairwise sample dependencies. When more than three samples are available, Eq. (1.4) can be extended further in a straightforward way. The same holds for categorical covariates, which can also be easily incorporated in Eq. (1.4). In this way the log-linear model provides a method to control for violations of both assumption 3 and 4 simultaneously. However, by extending the DSE model this way, a few additional issues are to be considered, which we will discuss in more detail in the next section, because they are related to the chapters in this dissertation.

1.2 Topics and main conclusions

This dissertation contains five main chapters. Four chapters are directly related to different methodological issues in MSE, and in a fifth chapter a large set of so-called nowcasting models that can also be applied to improve MSE estimates are discussed and compared. Each chapter in this dissertation was originally written as an independent article and therefore each chapter can be read independently. It is also important to note that the mathematical notation may differ to some degree between chapters because the notation is customised to each chapter.

The extension of DSE with additional samples and covariates to control for violations of DSE assumptions 3 and 4, is in different ways related to each of the chapters. A first issue that is introduced by this extension, is the increase of finite-sample bias that is caused by the decreasing value of the count variable n_{ab} with each split-up. In DSE the Chapman-estimator is available to improve the small-sample properties, but with multiple samples this solution is not available. This problem is discussed in Chapter 2 (see also Zult, van der Heijden, & Bakker, 2023). In this chapter a new estimator, the Chapman MSE-estimator, is proposed. This estimator is obtained by modifying the observed counts (e.g. n_{abc} in case of three samples) before estimation

is performed. This modification can be found in Eq. (2.32). This new estimator extends the Chapman-estimator in Eq. (1.2) towards multiple samples and categorical covariates, and outperforms other finite-sample bias reduced estimators that can be found in literature (i.e. Bailey, 1951; Evans & Bonett, 1994; Rivest & Lévesque, 2001; Cordeiro & McCullagh, 1991; Firth, 1993; Kosmidis, 2007; Kosmidis & Firth, 2011) in a series of simulation studies. This new Chapman MSE-estimator is also used to estimate the number of homeless people in The Netherlands, and in a comparison with the regular MSE estimates it shows substantially different results.

A second additional issue that is introduced by extending DSE with additional samples and/or categorical covariates is that it becomes unclear how to correct for bias due to a violation of assumption 2. Imperfect identification of population units leads to imperfect linkage and therefore to incorrect counts for n_{11}, n_{10} and n_{01} and consequently to a biased estimate for m_{00}. Ding and Fienberg (1994) and later Di Consiglio and Tuoto (2015) propose a DSE linkage-error corrected estimator. The basic idea behind this estimator is that if an audit sample is available that can be linked both probabilistically (see Fellegi & Sunter, 1969) and deterministically, the probabilities of missed links and false links can be estimated and these probabilities can be used to correct the population size estimate for linkage errors. However, it is unclear how this estimator can be extended towards multiple samples and/or covariates. This problem is dealt with in Chapter 3 and 4. In Chapter 3 a new linkage-error corrected DSE-estimator is proposed (see also de Wolf, van der Laan, & Zult, 2019) that is a generalisation of the estimator by Di Consiglio and Tuoto (2015), because in contrast to the estimator by Di Consiglio and Tuoto, it allows for samples of different sizes. Furthermore, in de Wolf et al. (2019) it is shown that this new linkage-error corrected estimator has an expectation that is equal to the expectation of the LP-estimator in Eq. (1.1). This result is used in Chapter 4, where the linkage-error corrected DSE-estimator in de Wolf et al. (2019) is further generalised towards MSE. This leads to a new linkage-error corrected MSE-estimator that we refer to as the weighted multiple - recapture (WMR) estimator. The basic idea behind the WMR-estimator is that the same audit sample that was assumed available with the other linkage-error corrected estimators, can also be used to create record-level weights. Summing up over these weights gives modified observed counts (e.g. n_{abc} in case of three samples) that can be used in the log-linear model to obtain a linkage-error corrected estimate for the size of the population. For two samples this new approach leads to the same linkage-error corrected estimator as derived in de Wolf et al. (2019), while it can be extended towards any number of samples in a fairly easy way. In a series of simulation studies it is shown that the WMR-estimator provides asymptotically unbiased estimators under different scenarios (see also Zult, de Wolf, Bakker, & van der Heijden, 2021).

Finally, a third additional issue that is introduced by extending DSE with additional samples, is that the use of additional samples may reduce the speed with which an estimate can be obtained, simply because some additional samples may become available later. This problem is discussed in Chapter 5 and 6. Both chapters discuss so-called nowcasting models, which are time series models that have the purpose of

obtaining an estimate for the current or most recent state of the time series. They are based on historic time series data and may be complemented with a subset of data that is available for the most recent period. Chapter 5 discusses a large set of nowcasting models that can be used to obtain a MSE-nowcast when a long enough time series of MSE-estimates would be available. Because such a (long enough) times series of MSE-estimates was not available, in Chapter 5 these nowcasting models are compared with the help of turnover data of companies in different economic sectors, for which much more and longer time series data is available (see Zult, Krieg, Schouten, Ouwehand, & van den Brakel, 2023). Finally, in Chapter 6 a new nowcasting model, designed specifically for MSE, is proposed. The idea behind this new model is to estimate Eq. (1.4) for the most recent period, by combining the most recent samples with samples from older periods with the help of the EM algorithm (see e.g. Dempster, Laird, & Rubin, 1977). This chapter discusses under which assumptions this model provides asymptotically unbiased population size estimates, and it is applied on the number of homeless people in The Netherlands, for which it shows reasonable results.

1.3 Open questions and further research

In Chapter 2 the properties of the new Chapman MSE-estimator are analysed both mathematically and in simulation studies. These analyses show that the Chapman MSE-estimator provides asymptotically unbiased population size estimates for a selection of MSE models, but neither approach provides complete proof in the sense that it shows that the Chapman MSE-estimator gives an asymptotically unbiased population size estimate for any possible MSE model. Such a complete proof is beyond the scope of this dissertation, but would be a valuable and welcome result of future research.

The excellent performance of the Chapman MSE-estimator compared to the standard MSE-estimator, as shown in Chapter 2, raises the question whether finite-sample bias correction should not become the standard approach in any work in the field of MSE research. Also, because it comes at almost no costs to researchers (see also Rainey & McCaskey, 2021). This will affect MSE estimates, and therefore also more general discussions, such as discussions on MSE model selection (e.g. in Silverman, 2020; Silverman, Chan, & Vincent, 2023; Binette & Steorts, 2022). For example, an interesting question would be if correcting for bias may also lead to the selection of different models.

The new linkage-error correction estimators in Chapter 3 and 4 are extensions of the linkage-error correction estimators by Ding and Fienberg (1994) and Di Consiglio and Tuoto (2015). One of the problems with all these estimators, is that they rely on the probabilistic linkage method by Fellegi and Sunter (1969) and the availability of so-called audit samples, which are sub-samples for which both probabilistic and perfect linkages based on a clerical review are available. Due to privacy regulations or simply because their construction is labour intensive, audit samples may not always

be available and therefore it is not always possible to obtain a linkage-error corrected estimate. This may change if an alternative probabilistic linkage procedure could be developed. The main purpose of the current probabilistic linkage method by Fellegi and Sunter (1969) is to link individual population units as accurate as possible, but for MSE the main concern is the accuracy of the aggregated counts n_{ab}, which is not necessarily the other side of the same coin. Therefore, if a probabilistic linkage method can be developed that optimises the accuracy of n_{ab} instead of the accuracy of the number of correct linkages, an audit sample may no longer be needed and obtaining a linkage-error corrected estimate more realistic. Whether and how such an alternative but valuable probabilistic linkage method would work is an open question and could be a topic of further research.

In Chapter 5 a large set of existing time series nowcasting models is tested and discussed. In Chapter 6 a new nowcasting model, designed specifically for MSE, is developed. The models from both these chapters can also be combined, which may further increase the accuracy of the resulting MSE estimates. However, to accurately test this combined approach, more and longer time series of MSE samples that were available for the work in Chapter 6 are required. This may therefore be a topic of further research.

Chapter 2

Bias correction in multiple systems estimation

If part of a population is hidden but two or more sources are available that each cover parts of this population, dual- or multiple system(s) estimation can be applied to estimate this population. For this it is common to use the log-linear model, estimated with maximum likelihood. These maximum likelihood estimates are based on a non-linear model and therefore suffer from finite-sample bias, which can be substantial in the case of small samples or a small population size. This problem was recognised by Chapman, who derived an estimator with good small sample properties in the case of two available sources. However, he did not derive an estimator for more than two sources. We propose an estimator that is an extension of Chapman's estimator to three or more sources and compare this estimator with other bias-reducing estimators in a simulation study. The proposed estimator performs well, and much better than the other estimators. A real data example on homelessness in the Netherlands shows that our proposed model can make a substantial difference.

A revised version of this chapter has been accepted, conditional on some small issues, for publication in the Journal of Official Statistics. A preliminary version is available at arXiv, i.e. Zult, D.B. (DZ), van der Heijden, P.G.M. (PvdH) and Bakker, B.F.M. (BB), 2023, Bias correction in multiple systems estimation, arXiv, https://doi.org/10.48550/arXiv.2311.01297. Author contributions: PvdH suggested the topic, DZ worked out the idea, did the analyses and wrote the text, PvdH and BB discussed and edited the text.

2.1 Introduction

A well-known statistical problem concerns the estimation of the size of a population that is only partly observed by different sources. By linking the records in the sources the number of units observed by at least one source is found, but the number of units that are missed by all sources is unknown. The standard method to estimate this hidden number is known as dual-system estimation (DSE) for two lists and multiple systems estimation (MSE) for more than two lists. Other names found in the literature are *capture-recapture*, *mark-recapture*, *multiple-recapture* and *multiple-record systems* estimation. A literature overview is provided by Chao, Tsay, Lin, and Chao (2001), who discuss these models in the context of human populations. An overview of its history and applications is provided by e.g. Cormack (1989); Bird and King (2018) and International Working Group for Disease Monitoring and Forecasting (1995a, IWGDMF).

DSE leans on a set of assumptions extensively described by, for example, Wolter (1986) and Zhang (2019). The IWGDMF summarize them as:

1. There is no change in the population during the investigation (the population is closed).

2. There is no loss of tags (individuals can be linked from capture to recapture).

3. For each sample, each individual has the same chance of being included in the sample.

4. The two samples are independent.

Earlier Seber (1982) and later Chao et al. (2001) and van der Heijden, Whittaker, Cruyff, Bakker, and van der Vliet (2012) showed that assumption 3 can be further relaxed, i.e., it is sufficient that each individual has the same chance of being included in only one of the samples, instead of both samples. When individuals have different inclusion probabilities, but covariates are available that are related to those, they can be included in the DSE or MSE model to control for them, (see e.g. Alho, 1990; Hook & Regal, 1995; Tilling & Sterne, 1999). In MSE assumption 4 can be relaxed, because samples are allowed to be dependent, and in practical situations this makes MSE much more realistic than DSE.

Under the appropriate assumptions and conditions, a maximum likelihood (ML) estimator can be derived for the hidden and total population size. However, in finite samples these ML-estimators are mean-biased, (see e.g. Chapman, 1951; Bailey, 1951; Rivest & Lévesque, 2001). This mean-bias can be shown for the ML-estimators directly, but also follows more generally from the fact that these estimators make use of a hierarchical log-linear model (Fienberg, 1972), which provides median-unbiased, but not mean-unbiased estimates, (see e.g. Miller, 1984; Hald, 1952, Ch. 7). This finite-sample mean-bias (from now on referred to as finite-sample bias or simply bias) can be substantial in the case of small samples (Rainey & McCaskey, 2021; Long, 1997,

p. 53-54), which in a DSE or MSE model can also occur if one source (or a combination of sources and categorical covariates) contains very few records (Tilling, 2001).

The role of finite-sample bias in the discussion on the robustness and accuracy of MSE-estimators is generally small. The focus is usually on other issues that lead to inaccurate estimates, such as failing model assumptions, (see e.g. Gerritse, van der Heijden, & Bakker, 2015; Zult et al., 2021) or model selection uncertainty (Silverman, 2020; Binette & Steorts, 2022). While it is true that these issues can potentially lead to large estimation bias, it is not clear how these issues are affected by finite-sample bias, simply because it is usually ignored. This is unfortunate, because correcting for finite-sample bias comes at almost no costs to researchers (Rainey & McCaskey, 2021), while, as we will see its impact can be substantial and therefore may affect conclusions.

The first to address the problem of finite-sample bias in DSE were Chapman (1951) and Bailey (1951). To reduce bias, they both proposed their own bias-reducing DSE estimator. Chapman showed that his estimator is "essentially unbiased" (Chapman, 1951, p. 145) and it became the most well-known of the two. Neither the Chapman nor Bailey estimator was extended towards MSE. The main contribution of this paper is the proposal of a Chapman MSE-estimator.

Our proposed Chapman MSE-estimator is not the first estimator that aims to reduce bias in the ML-estimator. Evans and Bonett (1994) and Rivest and Lévesque (2001) proposed population size estimators with the same goal. Others, such as Cordeiro and McCullagh (1991); Firth (1993); Kosmidis, Kenne Pagui, and Sartori (2020) and Kosmidis and Firth (2021), proposed bias-reduction methods for ML-estimators in log-linear models in general, which can be used in the context of MSE. In this paper we will compare the performance of these bias-reducing MSE-estimators with our Chapman MSE-estimator in simulation studies.

The paper is structured as follows. Section 2.2 discusses DSE and bias in DSE estimators. Section 2.3 discusses MSE and a derivation of the new Chapman MSE-estimator for *saturated* log-linear models, i.e., log-linear models where the number of independent parameters equals the number of counts. In Section 2.3.3 this new estimator is generalised towards a Chapman MSE-estimator that is also valid for *restricted* log-linear models. In Section 2.4 the new Chapman MSE-estimator is used to estimate the number of homeless people in The Netherlands. Section 2.5 discusses and concludes.

2.2 Dual-system estimation

This section discusses DSE. We first introduce notation, then Section 2.2.1 proceeds with the Lincoln-Peterson estimator and the log-linear model. Section 2.2.2 discusses the different distributional assumptions that underlie DSE and some of their implications. Section 2.2.3 introduces the problem of mean-bias and gives the bias-reducing DSE estimators proposed by Chapman (1951) and Bailey (1951). This section also

presents an alternative interpretation of the derivation of the Chapman-estimator that has the advantage that it allows the Chapman-estimator to be easily extended towards a similar estimator for multiple sources (which we will do in Section 2.3). Finally, in Section 2.2.4, bias-reducing DSE estimators are compared in a straightforward simulation study.

A description of the DSE problem starts from a population that consists of N unique units that are partly observed by two sources A and B, where the units are matched between sources. Each source is a random sample from the population, so in general not all N units are observed. Each unit has an inclusion pattern that tells us in which source(s) a unit was observed. This inclusion pattern is denoted as ab with $a, b = \{1, 0\}$, where $a = 1$ stands for "*in the first source*" and $a = 0$ for "*not in the first source*", and the same with b for the second source. This implies that the inclusion pattern 00 belongs to the unobserved units.

DSE uses the frequencies of occurrence of each inclusion pattern, which are simply the counts of the units with identical inclusion patterns. These counts are denoted as n_{ab}. A vector of the observed counts is denoted as $\mathbf{n}$, excluding the unobserved count n_{00}. When we sum over a or b, we replace that subscript by "+". Thus $n_{10} + n_{11} = n_{1+}$ is equal to the size of the first source, and n_{+1} to the size of the second source. The total number of observed units is denoted as n, which allows us to write $N = n + n_{00}$. n_{ab} and n_{00} are considered random variables with expectation m_{ab} and m_{00}. Estimates for N, m_{ab} and m_{00} are denoted by $\hat{N}^{\text{est}}$, $\hat{m}_{ab}^{\text{est}}$ and $\hat{m}_{00}^{\text{est}}$, where the superscript "est" indicates the estimator that was used. These bias-reducing estimators can be obtained by using adjusted counts, that we denote as n_{ab}^{est} or $\mathbf{n}^{\text{est}}$.

2.2.1 The Lincoln-Petersen estimator and the log-linear model

The first DSE model for population size estimation was proposed by Petersen (1896), and later Lincoln (1930). It is often referred to as the Lincoln-Petersen (LP) estimator. The LP-estimator can be derived from the assumption of independence between source A and B, which implies that the odds-ratio between source A and B, denoted by θ^{AB}, is

$$\theta^{AB} = \frac{m_{11}/m_{10}}{m_{01}/m_{00}} = 1, \tag{2.1}$$

which leads to

$$m_{00} = \frac{m_{10}m_{01}}{m_{11}}. \tag{2.2}$$

By plugging in ML estimates for m_{ab}, which are simply the observed values $n_{ab}^{\text{LP}} = n_{ab}$ (see e.g. Bishop et al., 1975), the LP-estimator for the missing cell is

$$\hat{m}_{00}^{\text{LP}} = \frac{n_{10}^{\text{LP}} n_{01}^{\text{LP}}}{n_{11}^{\text{LP}}} = \frac{n_{10}n_{01}}{n_{11}}, \tag{2.3}$$

and the population size estimate

$$\hat{N}^{\mathrm{LP}} = n + \hat{m}_{00}^{\mathrm{LP}} = \frac{n_{1+}n_{+1}}{n_{11}}. \tag{2.4}$$

The LP-estimator for the missing cell and for the population size are ML-estimators.

Fienberg (1972) shows that the LP-estimator can also be obtained from log-linear parameter estimates of the log-linear model

$$\log \mathrm{E}\left[\mathbf{n}|\mathbf{X}\right] = \mathbf{X}\boldsymbol{\lambda}, \tag{2.5}$$

with, for two sources, $\mathbf{n} = \mathbf{n}^{\mathrm{LP}} = (n_{11}, n_{01}, n_{01})^{\top}$, $\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ and $\boldsymbol{\lambda} = \left(\lambda, \lambda_a^A, \lambda_b^B\right)^{\top}$. λ is the intercept term, and λ_a^A and λ_b^B are the respective inclusion parameters for source A and B that are identified by setting $\lambda_0^A = \lambda_0^B = 0$. It is further assumed that Eq. (2.5) also holds for m_{00}. The parameters of a log-linear model are usually estimated with ML, which for Eq. (2.5) gives the ML estimates $\hat{\lambda}^{\mathrm{ML}}$, $\hat{\lambda}_a^{A,\mathrm{ML}}$ and $\hat{\lambda}_b^{B,\mathrm{ML}}$, which can be used to estimate m_{00}, i.e.:

$$\hat{m}_{00}^{\mathrm{ML}} = \exp \hat{\lambda}^{\mathrm{ML}}, \tag{2.6}$$

where $\hat{m}_{00}^{\mathrm{ML}}$ is equal to $\hat{m}_{00}^{\mathrm{LP}}$. It is well known that ML-estimators for log-linear models are biased (see e.g. Miller, 1984; Hald, 1952, Ch. 7), so this also holds for $\hat{m}_{00}^{\mathrm{LP}}$.

2.2.2 Distributional assumptions

Chapman (1951) and Bailey (1951) showed that the LP-estimator can be derived as an ML-estimator, assuming that n_{11} and n_{01} conditional on n_{1+} and N, follow a hypergeometric (Chapman) or binomial (Bailey) distribution. In the context of population size estimation, a hypergeometric distribution seems more fitting, because it assumes *sampling without replacement*, which matches the "no duplicates" assumption (i) of Zhang (2019). Bailey (1951, p. 294) was aware of this issue when he wrote "We shall assume that n_{+1} is sufficiently small compared with N for us to be able to ignore the complications of sampling without replacement". However, later Darroch (1958) argued that this choice is less obvious. He first showed that the LP-estimator can also be derived by assuming either

$$(n_{11}, n_{10}, n_{01}, n_{00}) \sim \text{Multinomial}\,(N, p_{11}, p_{10}, p_{01})$$

or

$$(n_{11}, n_{10}, n_{01}, n_{00}) \sim \text{Hypergeometric}\,(N, p_{11}, p_{10}, p_{01})$$

with $p_{ab} = m_{ab}/N$. Darroch (1958) discusses which of these distributions is the appropriate choice for any given experiment. He concludes that they lead to the same estimate $\hat{N}$ of N and the same asymptotic estimate of $\mathrm{Var}(\hat{N})$, so the difference is

notable only in higher moments. He further states that "In fact, if we had to generalise, we could say that the hypergeometric is likely to be appropriate when the main limiting factor on sample size is the trouble involved in marking animals and the multinomial when it is the difficulty in catching them.". This implies that, for instance, if a population is partly observed by lists of records that contain unique record ID-codes, the multinomial seems to be the most appropriate choice. Finally, Darroch (1958) concludes that the multinomial distribution is capable of generalisations that the hypergeometric is unable to accommodate, an advantage that we will use in this paper.

Later, Bishop et al. (1975, p. 446) showed that the assumption of a multinomial distribution can also be replaced by

$$n_{ab} \sim \text{Poisson}\,(m_{ab}),$$

with

$$n_{00} = N - n_{11} - n_{10} - n_{01},$$

without loss of generality. Both the multinomial and Poisson distribution have the practical advantage that they can deal with multiple sources more easily, but the Poisson distribution has a second advantage because it allows the simplification of some derivations due to $\text{Cov}\,(n_{ab}, n_{\neq ab}) = 0$ and $\text{Cov}\,(1/n_{ab}, n_{\neq ab}) = 0$.

2.2.3 Bias reduction in dual-system estimation

Chapman (1951) and Bailey (1951) were the first to be aware of the bias in the LP-estimator. This bias can be easily seen when we assume $n_{ab} \sim \text{Poisson}\,(m_{ab})$ and write the expectation of the LP-estimator as

$$\text{E}\left[\hat{m}_{00}^{\text{LP}}\right] = \text{E}\left[\frac{n_{10}n_{01}}{n_{11}}\right] = m_{10}m_{01}\,\text{E}\left[\frac{1}{n_{11}}\right], \tag{2.7}$$

which is not equal to $\frac{m_{10}m_{01}}{m_{11}}$ because $\text{E}\left[\frac{1}{n_{11}}\right] \neq \frac{1}{m_{11}}$. This shows that under a Poisson distribution, $\frac{1}{n_{11}}$ is the only source of bias in the LP-estimator, which was also noted by Rivest and Lévesque (2001).

Chapman and Bailey started with the hypergeometric and binomial distribution respectively and used different approximation approaches for the expectation of the ML-estimator to derive their bias-reducing estimators. Bailey used a second-order Taylor series approximation and concludes that

$$\hat{m}_{00}^{\text{Bailey}} = \frac{n_{10}^{\text{Bailey}}\, n_{01}^{\text{Bailey}}}{n_{11}^{\text{Bailey}}} = \frac{n_{10}(n_{01}-1)}{(n_{11}+1)} \tag{2.8}$$

and

$$\hat{N}^{\text{Bailey}} = \frac{n_{1+}(n_{+1}+1)}{(n_{11}+1)}. \tag{2.9}$$

are bias-reducing estimators for m_{00} and N respectively (Bailey, 1951, p. 295).

Chapman uses a different approach that is recommended by Stephan (1945). Instead of a Taylor approximation, Stephan recommends writing $\mathrm{E}\left[\frac{1}{x}\right]$, with x a binomial random variable, as a series of inverse factorials, as one needs quite a few terms before a Taylor series becomes reasonably accurate (Stephan, 1945, p. 52). This increased rate of convergence of Stephan's inverse factorial approximation in the case of $\mathrm{E}\left[\frac{1}{x}\right]$ and $n_{11} \sim \text{Poisson}(m_{11})$, is illustrated with a straightforward simulation study presented in Appendix 2.6.1. Chapman uses Stephan's inverse factorial approximation to derive a bias-reducing expression for $\frac{n_{10}n_{01}}{n_{11}}$ and concludes that a bias-reducing estimator for m_{00} is

$$\hat{m}_{00}^{\text{Chap}} = \frac{n_{10}^{\text{Chap}} n_{01}^{\text{Chap}}}{n_{11}^{\text{Chap}}} = \frac{n_{10}n_{01}}{(n_{11}+1)}, \tag{2.10}$$

and for N

$$\hat{N}^{\text{Chap}} = \frac{(n_{1+}+1)(n_{+1}+1)}{(n_{11}+1)} - 1. \tag{2.11}$$

A Bailey or Chapman estimate can also be obtained from the log-linear model in Eq. (2.5), if instead of $\mathbf{n}^{\text{LP}} = (n_{11}, n_{10}, n_{01})^\top$, respectively $\mathbf{n}^{\text{Bailey}} = (n_{11}+1, n_{10}, n_{01}-1)^\top$ or $\mathbf{n}^{\text{Chap}} = (n_{11}+1, n_{10}, n_{01})^\top$ is used.

The Chapman- and Bailey-estimator differ only slightly, but the Chapman-estimator became the standard bias-reducing estimator in the DSE literature. A good reason is that Chapman (1951, p. 146) further shows that if $\frac{n_{1+}n_{+1}}{N} > \log\left(\frac{N}{\epsilon}\right)$ holds, then

$$\left|\mathrm{E}\left[\hat{N}^{\text{Chap}}\right] - N\right| < \frac{\epsilon}{100}N,$$

with ϵ some arbitrary small positive number (Cramer, 1922, p. 502), also holds. This means that if the two sources are large enough compared to N, the bias in $\hat{N}^{\text{Chap}}$ is less than ϵ percent of N and so Chapman refers to his estimator as "essentially unbiased". Therefore we refer to the Chapman-estimator not only as a bias-reducing, but also as a bias-*corrected* estimator. Chapman (1951, p. 146) finally notes that the Chapman-estimator requires

$$\frac{n_{1+}n_{+1}}{N} > \log N, \tag{2.12}$$

to hold. This inequality is derived from setting $\left|\mathrm{E}\left[\hat{N}^{\text{Chap}}\right] - N\right| \leq 1$ and can be considered a regularity condition for the Chapman-estimator. If this regularity condition is not met, $\hat{N}^{\text{Chap}}$ may suffer from considerable (negative) bias, as we will illustrate later in scenario 7 in the simulation study in Table 2.1. Later, Wittes (1972) showed that the Chapman-estimator is unbiased if $n_{1+} + n_{+1} > N$.

Chapman derived his estimator for the hypergeometric distribution, but it can also be developed with second-order Taylor approximations for the multinomial and

Poisson distributions, which are derived in Appendix 2.6.2. This derivation suggests that the Chapman-estimator is also valid under a multinomial or Poisson distribution. This is useful when we extend the Chapman-estimator to multiple sources in Section 2.3.2. Combining the Chapman-estimator with the results in Appendix 2.6.1 and 2.6.2 implies that if $n_{ab} \sim \text{Poisson}(m_{ab})$ we can write

$$\frac{1}{m_{ab}} \approx \text{E}\left[\frac{1}{n_{ab}+1}\right]. \tag{2.13}$$

This equation will allow us to easily extend the Chapman MSE-estimator towards multiple sources in Section 2.3.2.

Bailey did not extend his estimator to more than two sources. Chapman (1952) did, but he only considered the case where a unit was tagged in an earlier source or not, and did not consider dependence between pairs of sources. Dependence between sources is further discussed in Section 2.3.1. Others, like Cordeiro and McCullagh (1991); Firth (1993); Evans and Bonett (1994); Rivest and Lévesque (2001) and Kosmidis et al. (2020) have proposed bias-reducing estimators for log-linear models in general and therefore do take dependence between sources into account. These models are discussed in more detail in Section 2.3.1.1. However, we will include these bias-reducing estimators in the simple DSE simulation study presented in Section 2.2.4.

2.2.4 Dual-system estimation simulation study

In this section we compare the LP-, Bailey-, Cordeiro, Firth-, Kosmidis, Evans and Bonette (EB)-, Rivest and Lévesque (RL)- and Chapman-estimator in a DSE setting. The LP-, Bailey- and Chapman-estimator can only be used in DSE and were discussed in the previous sections. The Cordeiro, Firth-, Kosmidis, EB- and RL-estimator can be applied in both DSE and MSE and will be discussed in Section 2.3.1.1. The method that is used to generate contingency tables is discussed in Hammond, van der Heijden, and Smith (2024). It allows the generation of contingency tables with a log-linear model that has prespecified inclusion probabilities p_A and p_B and odds ratio(s). The resulting n_{ab} are generated from a multinomial distribution. This is particularly useful in the next section in which we consider more than two sources, and pairs of sources that are dependent.

A minor but important simulation issue is the regularity condition in Eq. (2.12), or the issue of what Otis, Burnham, White, and Anderson (1978, p. 125) refer to as "failures". This implies that the relation between n_{1+}, n_{+1} and N must be set such that they comply with Eq. (2.12). A simple example of a failure is when, in DSE, n_{11} equals zero, which leads to $\hat{N}^{\text{LP}} = \infty$. Otis et al. (1978) recommend replacing such a replication with a new replication, an advice that was followed in Evans and Bonett (1994). However, replacing failure replications, that correspond to large population size estimates, with new replications, introduces selection bias in the sense that, when $\hat{N}^{\text{est}}$ is an unbiased estimator for N, the mean of these estimates $\bar{N}^{\text{est}} = \sum_{r=1}^{R} \hat{N}_r^{\text{est}}/R$ with R the number of replications, departs from N. Therefore, to obtain accurate

mean estimates that allow a fair comparison of bias between the different estimators, we choose the combined N, p_A and p_B such that for the scenarios $S = 1, 2, \ldots, 6$ (see Table 2.1) the probability of failures becomes close to zero. Nonetheless a failure occurred once for scenario 1. These settings also imply that the regularity condition in Eq. (2.12) holds by a substantial margin. To see how estimators are affected when the regularity condition is violated, we have added a 7^{th} scenario under which the regularity condition does not hold, i.e. $n_{1+} = n_{+1} = 15$, so $n_{1+}n_{+1}/N = 225/100 < \log 100$.

The scenario parameters are shown in the columns N, p_A and p_B of Table 2.1 below. The different estimators that are compared are shown in the columns that follow. In the context of DSE some estimators are equivalent and their results are displayed in a single column. This holds for $\hat{N}^{\text{EB}}$, $\hat{N}^{\text{Cordeiro}}$, $\hat{N}^{\text{Firth}}$ and $\hat{N}^{\text{Kosmidis}}$ (denoted as $\hat{N}^{\text{EB/CFK}}$), and for $\hat{N}^{\text{Chap}}$ and $\hat{N}^{\text{RL}}$ (denoted as $\hat{N}^{\text{Chap/RL}}$).

Table 2.1: Simulation study with 20,000 replications for seven DSE scenarios.

S	N	p_A	p_B	$\bar{n}$	$\bar{N}^{\text{LP}}$	$\bar{N}^{\text{Bailey}}$	$\bar{N}^{\text{EB/CFK}}$	$\bar{N}^{\text{Chap/RL}}$
1	100	0.5	0.2	60.0	$105.3^{***\dagger}$	96.1^{***}	105.2^{***}	100.1
2	100	0.35	0.3	54.5	106.0^{***}	98.0^{***}	105.3^{***}	100.4^{*}
3	500	0.4	0.15	244.9	508.3^{***}	493.6^{***}	507.4^{***}	499.2
4	500	0.25	0.2	200.1	512.4^{***}	495.4^{***}	509.3^{***}	499.4
5	10,000	0.3	0.1	3,699.2	$10{,}018.0^{***}$	$9{,}987.9^{***}$	$10{,}013.1^{***}$	9,996.9
6	10,000	0.25	0.15	3,624.9	$10{,}016.7^{***}$	$9{,}993.9^{*}$	$10{,}012.5^{***}$	9,999.6
7	100	0.15	0.15	27.8	$146.2^{***\dagger}$	87.2^{***}	128.0^{***}	92.3^{***}

$\bar{n}$ gives the mean number of observed units n over all replications. The superscripts *, ** and *** indicate that we can reject $\hat{N}^{\text{est}} = N$ with a two-sided t-test with p-values = 0.05, 0.01 and 0.001 respectively. A † as superscript indicates that extremely high estimates due to failures were replaced with the corresponding $\hat{N}^{\text{EB}}$ for that replication.

The *s in the column of $\bar{N}^{\text{Chap/RL}}$ indicate that for p-value = 0.05, in five out of the six regular scenarios, the hypothesis $N = \hat{N}^{\text{Chap/RL}}$ cannot be rejected. For p-value = 0.01 this holds for all six regular scenarios. The same does not hold for the other estimators, for which the mean over all replications, in most cases, significantly differs from N for p-value = 0.001, and for all cases for p-value = 0.05. For all scenarios the bias in $\hat{N}^{\text{Chap/RL}}$ is smaller than the bias in the other estimators. This shows that in DSE, the Chapman- and RL-estimator are superior to the other estimators. If Chapman's regularity condition in Eq. (2.12) is not met, as in scenario 7, all estimators are considerably biased.

The standard error (SE) and root mean squared errors (RMSEs) that correspond to each estimator and scenario in Table 2.1 can be found in Table 2.8 in Appendix 2.6.3.1. This table shows that the SEs and RMSEs of the Bailey- and Chapman/RL-estimator are smaller than the RMSEs of the EB/CFK-estimator, which in turn are smaller than the RMSEs of the ML-estimator.

2.3 Multiple systems estimation

This section discusses multiple systems estimation (MSE). First it introduces some notation additional to the DSE notation introduced in Section 2.2. Next, Section 2.3.1 proceeds with some MSE preliminaries and bias-reducing MSE-estimators. In Section 2.3.2 we derive a new bias-corrected estimator that can be considered an extension of the Chapman-estimator towards MSE under saturated models. In Section 2.3.3 the Chapman MSE-estimator is further generalised towards all log-linear models, both saturated and restricted.

MSE considers the case where a population that consists of N unique units is partly observed by a set of k sources, indicated by $A, B, C, ..., K$. For ease of notation we will, where possible, discuss MSE from the perspective of three sources, because it can often be generalised to k sources in a straightforward way. For three sources, the inclusion pattern is denoted as abc with $a, b, c = 1, 0$, with the same meaning as ab in DSE notation. For k sources the inclusion pattern is $ab \dots k$. We introduce notation that allows us to distinguish between the sets of unit counts that are observed an *even* and *odd* number of times, that we denote by $\mathbf{n}_{\text{even}}$ (or $\mathbf{m}_{\text{even}}$, $\hat{\mathbf{m}}_{\text{even}}$) and $\mathbf{n}_{\text{odd}}$ (or $\mathbf{m}_{\text{odd}}$, $\hat{\mathbf{m}}_{\text{odd}}$). For three sources this gives $\mathbf{n}_{\text{odd}} = (n_{111}, n_{100}, n_{010}, n_{001})$ and $\mathbf{n}_{\text{even}} = (n_{110}, n_{101}, n_{011})$.

In contrast to DSE, in MSE the log-linear model can take different forms. Therefore, the superscript in $\hat{N}^{\text{est}}$, n_{ab}^{est} and $\hat{m}_{ab}^{\text{est}}$ is extended to $\hat{N}^{\text{est, LLM}}$, $n_{abc}^{\text{est, LLM}}$ and $\hat{m}_{abc}^{\text{est, LLM}}$, where "est, LLM" specifies not only the chosen estimator but also the chosen log-linear model.

2.3.1 Preliminaries

The first to consider more than two sources was Schnabel (1938). After this the use of multiple sources became more common and estimators were introduced that made use of different distributional assumptions. For instance, Chapman (1954); Darroch (1958) and Cormack and Jupp (1991) assumed every element in n_{abc} to be an independent realisation from a Poisson distribution. This is a reasonable assumption when n_{abc} are relatively small compared to N, but when this is not true, one should take into account that each m_{abc} has an upper-bound of N. Adding this restriction to the Poisson distribution assumption is equivalent to assuming that the joint set of n_{abc} has a multinomial distribution with expectations m_{abc} for which $m_{000} + \sum_{abc} m_{abc} = N$, (see e.g. Sanathanan, 1972; Bishop et al., 1975; Wolter, 1986; Darroch, Fienberg, Glonek, & Junker, 1993).

In the case of three sources the independence assumption that holds in DSE is relaxed and it is sufficient to assume that two conditional odds-ratios given the levels of the third source are equal. For example, for the two odds-ratios of source A and B

given source C

$$\frac{m_{110}/m_{100}}{m_{010}/m_{000}} = \frac{m_{111}/m_{101}}{m_{011}/m_{001}}, \tag{2.14}$$

which gives

$$m_{000} = \frac{m_{111}m_{100}m_{010}m_{001}}{m_{110}m_{101}m_{011}}. \tag{2.15}$$

A general expression is provided by Fienberg (1972), who states that for k sources, $m_{00\ldots0}$ can be written as

$$m_{00\ldots0} = \frac{\prod \mathbf{m}_{\text{odd}}}{\prod \mathbf{m}_{\text{even}}}. \tag{2.16}$$

For three sources, the saturated (SAT) log-linear model for the seven observed counts becomes

$$\text{SAT: } \log m_{abc} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}, \tag{2.17}$$

where the parameters are identified by setting them to zero if one or more of the subscripts are 0. In comparison to DSE, in the saturated log-linear model the independence assumption is replaced by the assumption of no three-factor interaction, i.e $\lambda_{abc}^{ABC} = 0$. The interaction parameters λ_{ab}^{AB}, λ_{ac}^{AC} and λ_{bc}^{BC} allow for interactions between pairs of sources, and thus the model is less restrictive than the DSE model and hence more realistic in applications.

For three sources, the saturated model is not the only log-linear model that can be used. If the parameters of one or more pairs of sources are set to zero (e.g. $\lambda_{ab}^{AB} = 0$), we have a *restricted* log-linear model. An advantage of further restricted models is that the resulting estimates have smaller variance than estimates from less restricted models (Bishop et al., 1975, p. 242). A disadvantage is that they give biased estimates if the assumed restriction does not hold. We discuss restricted models in more detail because, as will be shown in Section 2.3.3, the chosen model specification affects the bias-corrected estimator. Fienberg (1972) and Bishop et al. (1975) discuss the three possible alternative log-linear model formulations for three sources where all direct inclusion parameters λ_a, λ_b and λ_c are included. Starting from the saturated log-linear model SAT in Eq. (2.17), they discuss the two-pair dependence (2PD), the one-pair dependence (1PD), and independence (IND) model. Examples of 2PD and 1PD are

$$\text{2PD} : \log m_{abc} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{bc}^{BC}, \tag{2.18}$$

$$\text{1PD} : \log m_{abc} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} \tag{2.19}$$

and

$$\text{IND} : \log m_{abc} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C. \tag{2.20}$$

It suits our purpose to write these models as in Eq. (2.5). They all use $\mathbf{n} = \mathbf{n}^{\text{ML}} = (n_{111}, n_{110}, n_{101}, n_{011}, n_{100}, n_{010}, n_{001})$, but differ with respect to $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{LLM}}$ and $\mathbf{X} = \mathbf{X}^{\text{LLM}}$. $\boldsymbol{\lambda}^{\text{LLM}}$ simply consists of the λ's in the corresponding LLM and $\mathbf{X}^{\text{LLM}}$ becomes X^{SAT}_{abc}, X^{2PD}_{abc}, X^{1PD}_{abc} or X^{IND}_{abc} written as

$$\begin{pmatrix} 1&1&1&1&1&1&1\\ 1&1&1&0&1&0&0\\ 1&1&0&1&0&1&0\\ 1&0&1&1&0&0&1\\ 1&1&0&0&0&0&0\\ 1&0&1&0&0&0&0\\ 1&0&0&1&0&0&0 \end{pmatrix}, \begin{pmatrix} 1&1&1&1&1&1\\ 1&1&1&0&1&0\\ 1&1&0&1&0&0\\ 1&0&1&1&0&1\\ 1&1&0&0&0&0\\ 1&0&1&0&0&0\\ 1&0&0&1&0&0 \end{pmatrix}, \begin{pmatrix} 1&1&1&1&1\\ 1&1&1&0&1\\ 1&1&0&1&0\\ 1&0&1&1&0\\ 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0 \end{pmatrix} \text{ or } \begin{pmatrix} 1&1&1&1\\ 1&1&1&0\\ 1&1&0&1\\ 1&0&1&1\\ 1&1&0&0\\ 1&0&1&0\\ 1&0&0&1 \end{pmatrix},$$

respectively. Estimating one of these models with ML gives the fitted-values $\hat{m}^{\text{ML, LLM}}_{abc}$.

A general expression for $m^{\text{ML, LLM}}_{000}$ is obtained by replacing the m_{abc} in Eq. (2.15) with the ML estimates $\hat{m}^{\text{ML, LLM}}_{abc}$. For LLM = SAT this gives $m^{\text{ML, LLM}}_{abc} = m^{\text{SAT, ML}}_{abc} = n_{abc}$ and thus

$$\hat{m}^{\text{SAT, ML}}_{000} = \frac{n_{111}n_{100}n_{010}n_{001}}{n_{110}n_{101}n_{011}}. \tag{2.21}$$

For model 2PD and 1PD, Fienberg (1972, p. 596) shows this expression can be further simplified, i.e.,

$$\hat{m}^{\text{2PD, ML}}_{000} = \frac{n_{100}n_{001}}{n_{101}} \tag{2.22}$$

and

$$\hat{m}^{\text{1PD, ML}}_{000} = \frac{n_{001}n_{++0}}{n_{111} + n_{101} + n_{011}}. \tag{2.23}$$

For model IND, such a closed form solution does not exist. However, for IND an estimate can be obtained by replacing the m_{abc} in Eq. (2.15) with the fitted-values $\hat{m}^{\text{IND, ML}}_{abc}$. Fienberg (1972, p. 597) shows that this is in fact an approach that can be generally used, where Eq. (2.16) gives an estimate of the missing cell for any log-linear model with any number of sources. Note that the LP-estimator in Eq. (2.3) can be considered a special of (2.16), with $k = 2$ and LLM = SAT.

2.3.1.1 Bias reduction in multiple systems estimation

Bias in MSE models was considered by Evans and Bonett (1994) and Rivest and Lévesque (2001), who propose MSE-estimators with improved finite-sample properties. More general bias-reducing estimators for ML estimates in generalised linear models, such as the log-linear model, are introduced in Cordeiro and McCullagh (1991); Firth (1993). These models are also known as generalised linear models (GLMs) using adjusted score functions see also Kosmidis et al. (2020) for further

discussion. A DSE example of this approach can be found in Section 2.2.3, where replacing $\mathbf{n}$ with the adjusted $\mathbf{n}^{\text{Chap}}$ or $\mathbf{n}^{\text{Bailey}}$ led to the bias-reducing DSE estimators by Chapman and Bailey. These more general estimators aim to derive the first-order bias term from the ML-estimators, (see e.g. Sowden, 1972). Cordeiro and McCullagh (1991) derive a first-order bias expression for generalised linear models, such as the log-linear model, from the bias expression by McCullagh and Nelder (1989, 5.13, p. 456), and use it to reduce bias. Firth (1993) showed that this reduction can also be achieved by iteratively modifying the score function, which in a Poisson log-linear model implies iteratively modifying the dependent count variable. For simple MSE models this iterative modification converges to the modification used in the more straightforward bias-reducing MSE-estimator by Evans and Bonett (1994). Rivest and Lévesque (2001) also derive a modification scheme to remove first-order bias for the MSE models proposed by Otis et al. (1978), which correspond to a selected set of log-linear models (Chao, 2001). For two samples the estimator by Rivest and Lévesque (2001) is equal to the Chapman-estimator and for more than two samples they show that their estimator outperforms the estimator by Evans and Bonett (1994) in a Monte Carlo experiment.

The simplest bias-reducing MSE-estimator is proposed by Evans and Bonett (1994, EB), which we denote as $\hat{N}^{\text{EB,LLM}}$. They propose to use the adjusted $\mathbf{n}^{\text{EB}} = \mathbf{n} + 0.5^{(k-1)}$ in (2.5) instead of $\mathbf{n}$. This is the result of a compromise between Berkson (1955) and Plackett (1981). In a log-linear regression model, Berkson proposes to replace values in n_{abc} that are equal to zero with $0.5^{(k-1)}$, and Plackett (1981), who states that if $n_{abc} \sim \text{Poisson}(m_{abc})$, then $\log(n_{abc} + 0.5)$ instead of $\log(n_{abc})$ is a more accurate estimate for $\log(m_{abc})$.

Another bias-reducing estimator that was developed specifically for MSE was proposed by Rivest and Lévesque (2001, RL). They propose a bias reduction method that can be used to reduce bias in a set of MSE-estimators proposed by Otis et al. (1978) in the context of wildlife populations. Unfortunately, with the exception of the independence model, which corresponds to the M_t model, the other models by Otis et al. do not correspond exactly to Eq. (2.17) - (2.19). For the SAT, 2PD and 1PD model, we consider the adjusted counts that belong to model M_{th} as the most appropriate choice, because it is most similar. See Evans, Bonett, and McDonald (1994) for further discussion on this topic. The bias reduction by Rivest and Lévesque is derived from a standard result by McCullagh and Nelder (1989), about the bias in estimators in generalised linear models. McCullagh and Nelder derive an asymptotic bias expression for estimates based on models with canonical link functions, such as the log-linear model. For two sources, the resulting RL-estimator is equal to the Chapman-estimator, as was seen in Table 2.1. We denote the RL-estimator as $\hat{N}^{\text{RL, LLM}}$ and their adjusted count as $\mathbf{n}^{\text{RL, LLM}}$, with $\mathbf{n}^{\text{RL, IND}} = \mathbf{n}^{\text{RL, } M_t}$ and $\mathbf{n}^{\text{RL, SAT/2PD/1PD}} = \mathbf{n}^{\text{RL, } M_{th}}$. For three sources they become (Rivest & Lévesque, 2001, p. 562):

$$\mathbf{n}^{\text{RL, } M_t} = (n_{111}, n_{110} + \frac{1}{3}, n_{101} + \frac{1}{3}, n_{011} + \frac{1}{3}, n_{100} + \frac{1}{6}, n_{010} + \frac{1}{6}, n_{001} + \frac{1}{6})^\top \tag{2.24}$$

and

$$\mathbf{n}^{\text{RL, } M_{th}} = (n_{111}, n_{110} + \frac{1}{3}, n_{101} + \frac{1}{3}, n_{011} + \frac{1}{3}, n_{100}, n_{010}, n_{001})^\top. \tag{2.25}$$

Rivest and Lévesque (2001) show in a simulation study that their estimator outperforms the EB-estimator in terms of bias reduction.

Bias reduction in MSE by means of the modified-score functions approach (Firth, 1993) relies on the same standard result about the bias of estimators in GLMs by McCullagh and Nelder (1989) as was used by Rivest and Lévesque (2001). It was used by Cordeiro and McCullagh (1991); Firth (1993); Kosmidis (2007); Kosmidis and Firth (2011) and others to reduce bias in parameter estimates in log-linear models. Cordeiro and McCullagh; Firth and Kosmidis and Firth give three different bias-reducing parameter estimates $\hat{\lambda}^{\text{est}}$ for the λ in Eq. (2.5), which correspond to three different bias-reducing estimators $\hat{m}^{\text{est}}_{000} = \exp \hat{\lambda}^{\text{est}}$. The description of the details on these estimators are beyond the scope of this paper, but they are provided in Kosmidis (2014); Kosmidis et al. (2020) and Kosmidis and Kenne Pagui (2023). In this paper we limit ourselves to noting that in the DSE and MSE simulation studies presented in this paper we found negligible differences between them, and therefore we denote them as the single estimator $\hat{N}^{\text{CFK, LLM}}$.

In the next section we extend the Chapman-estimator towards multiple sources, which results in a Chapman MSE-estimator that differs from the estimators discussed in this section, both for the saturated and more restricted log-linear models.

2.3.2 The Chapman MSE-estimator for saturated models

To derive a Chapman MSE-estimator, we start with the result of Bishop et al. (1975, p. 446), who showed that ML-estimators for m_{abc} are equivalent under the assumption that n_{abc} follows either a Poisson or multinomial distribution, provided that $\sum_{abc} m_{abc} + m_{000} = N$. Combining the implications of the Chapman-estimator as discussed in Section 2.2.3 with the MSE models discussed in Section 2.3.1 under the assumption of a Poisson distribution allows us to derive a bias-corrected MSE-estimator in a straightforward way. The Poisson distribution implies that $\text{Cov}(n_{abc}, n_{\neq abc}) = 0$ and $\text{Cov}(1/(n_{abc}+1), n_{\neq abc}) = \text{Cov}(1/n_{abc}, n_{\neq abc}) = 0$, when we combine this with Eq. (2.13) and (2.16) this gives

$$m_{00\ldots0} = \frac{\prod \mathbf{m}_{\text{odd}}}{\prod \mathbf{m}_{\text{even}}} \approx \prod \text{E}\left[\mathbf{n}_{\text{odd}}\right] \prod \text{E}\left[\frac{1}{(\mathbf{n}_{\text{even}}+1)}\right] = \text{E}\left[\frac{\prod \mathbf{n}_{\text{odd}}}{\prod (\mathbf{n}_{\text{even}}+1)}\right], \tag{2.26}$$

which suggests

$$\hat{m}^{\text{SAT, Chap MSE}}_{000} = \frac{n_{111} n_{100} n_{010} n_{001}}{(n_{110}+1)(n_{101}+1)(n_{011}+1)} \tag{2.27}$$

as a bias-corrected estimator for three sources, and

$$\hat{m}^{\text{SAT, Chap MSE}}_{00\ldots0} = \frac{\prod \mathbf{n}_{\text{odd}}}{\prod (\mathbf{n}_{\text{even}}+1)} \tag{2.28}$$

for any number of sources.

When we compare the Chapman MSE-estimator in Eq. (2.27) with the RL-estimator in Eq. (2.25), it becomes clear that the equality between both estimators in DSE does not hold for MSE. We further note that Chapman MSE estimates can also be obtained with the Poisson regression model as defined in Eq. (2.5), by using the modified counts $n_{abc}^{\text{SAT, Chap MSE}}$ instead of $\mathbf{n}$.

2.3.2.1 Simulation study with saturated models

In Section 2.2.4 we have seen that the Chapman- and RL-estimator are equivalent and less biased than the alternative DSE estimators. This equivalence is unlikely to hold in MSE, because they are no longer the same estimators. Here we compare them in a simulation study, together with other bias-reducing MSE-estimators. We consider fourteen scenarios. The scenarios in Table 2.2 differ with respect to the size of the population N, the number of sources k and log-linear model specifications (i.e. different values for p_A, p_B, p_C, p_D, θ^{AB}, θ^{AC}, θ^{AD}, θ^{BC} and θ^{CD}, see Hammond et al. (2024) for further details. The odds-ratios are chosen such that scenario $1-3$ and $13-14$ concern $\text{LLM}^{S_i} = \text{IND}$, scenario $4-6$ concern $\text{LLM}^{S_i} = \text{1PD}$, scenario $7-9$ concern $\text{LLM}^{S_i} = \text{2PD}$, scenario $10-12$ concern $\text{LLM}^{S_i} = \text{SAT}$ and finally scenario 15 concerns $\text{LLM}^{S_i} = \text{4PD}$ (i.e. four pairs of dependent sources), with LLM^{S_i} the log-linear model used to generate the contingency table. The different parameters are chosen such that the probability of failures in each scenario is small.

Table 2.2: MSE simulation scenarios

S_i	N	s	p_A	p_B	p_C	p_D	θ_{AB}	θ_{AC}	θ_{BC}	θ_{AD}	θ_{BD}	θ_{CD}	Corresponding model
1	100	3	0.5	0.4	0.3		1	1	1				Independence
2	500	3	0.4	0.3	0.2		1	1	1				Independence
3	10,000	3	0.35	0.3	0.25		1	1	1				Independence
4	100	3	0.5	0.4	0.3		1.5	1	1				One-pair dependence
5	500	3	0.4	0.3	0.2		1.5	1	1				One-pair dependence
6	10,000	3	0.35	0.3	0.25		1.5	1	1				One-pair dependence
7	100	3	0.5	0.4	0.3		1.5	1	0.5				Two-pair dependence
8	500	3	0.4	0.3	0.2		1.5	1	0.5				Two-pair dependence
9	10,000	3	0.35	0.3	0.25		1.5	1	0.5				Two-pair dependence
10	100	3	0.5	0.4	0.3		1.5	0.75	0.5				Saturated
11	500	3	0.4	0.3	0.2		1.5	0.75	0.5				Saturated
12	10,000	3	0.35	0.3	0.25		1.5	0.75	0.5				Saturated
13	1,000	4	0.35	0.3	0.25	0.2	1	1	1	1	1	1	Independence[1]
14	20,000	4	0.25	0.2	0.15	0.1	1	1	1	1	1	1	Independence[1]
15	20,000	4	0.25	0.2	0.15	0.1	1.5	1	0.75	1.5	1	0.5	Four-pair dependence[1]

[1] The three-way interaction parameters θ_{ABC}, θ_{ACD} and θ_{BCD} are set to 1.

The estimates presented in Table 2.3 below are based on the saturated model. This means that for all scenarios, except scenario $10-12$, the model is overspecified. Over-

Table 2.3: Simulation study with assumed saturated log-linear models, with 100,000 replications for MSE scenarios 1 – 15, for MSE scenario 1 – 15 in Table 2.2.

S	N	$\bar{n}$	$\bar{N}^{\text{SAT, ML}}$	$\bar{N}^{\text{SAT, EB}}$	$\bar{N}^{\text{SAT, CFK}}$	$\bar{N}^{\text{SAT, RL}}$	$\bar{N}^{\text{SAT, Chap MSE}}$
1	100	79.0	$113.2^{***\dagger}$	112.3^{***}	110.9^{***}	103.3^{***}	100.0
2	500	332.0	521.8^{***}	522.1^{***}	522.4^{***}	507.0^{***}	500.2
3	10,000	6,587.4	$10{,}016.0^{***}$	$10{,}016.8^{***}$	$10{,}017.6^{***}$	$10{,}004.6^{***}$	9,999.0
4	100	77.3	$116.9^{***\dagger}$	115.2^{***}	112.7^{***}	104.0^{***}	99.9
5	500	323.8	525.3^{***}	524.5^{***}	523.9^{***}	508.4^{***}	500.8^{***}
6	10,000	6,439.5	$10{,}017.8^{***}$	$10{,}018.0^{***}$	$10{,}018.3^{***}$	$10{,}005.6^{***}$	9,999.4
7	100	79.1	$120.6^{***\dagger}$	118.7^{***}	113.6^{***}	104.0^{***}	99.6^{***}
8	500	330.5	532.2^{***}	530.8^{***}	529.6^{***}	510.0^{***}	500.3
9	10,000	6,608.9	$10{,}020.9^{***}$	$10{,}021.6^{***}$	$10{,}022.4^{***}$	$10{,}007.3^{***}$	10,000.6
10	100	80.0	$118.2^{***\dagger}$	116.6^{***}	112.8^{***}	103.8^{***}	99.7^{***}
11	500	334.1	$529.5^{***\dagger}$	529.4^{***}	529.3^{***}	508.9^{***}	499.9
12	10,000	6,690.3	$10{,}021.2^{***}$	$10{,}022.6^{***}$	$10{,}023.9^{***}$	$10{,}008.0^{***}$	10,001.5
13	1,000	727.0	$1{,}205.3^{***\dagger}$	$1{,}196.4^{***}$	$1{,}134.2^{***}$	$1{,}183.7^{***}$	996.8^{***}
14	20,000	10,819.7	$20{,}904.5^{***}$	$20{,}859.1^{***}$	$20{,}730.0^{***}$	$20{,}846.6^{***}$	20,005.8
15	20,000	10,677.5	$21{,}333.1^{***}$	$21{,}281.8^{***}$	$21{,}143.0^{***}$	$21{,}255.0^{***}$	$20{,}033.2^{**}$

$\bar{n}$ gives the mean number of observed units n over all replications. The superscripts *, ** and *** indicate that we can reject $\hat{N}^{\text{est}} = N$ with a two-sided t-test with p-values = 0.05, 0.01 and 0.001 respectively. A † as superscript indicates that extremely high estimates due to failures were replaced with the corresponding $\hat{N}^{\text{SAT, EB}}$ for that replication.

specification only affects the variance and not the mean of an estimator, so it does not lead to the introduction of bias, although it may increase the bias when it is present, which is discussed in more detail below Table 2.5. In contrast to the DSE simulation study in Section 2.2.4, it was not possible to exclude failures in all scenarios, in particularly for $N = 100$. In those cases the estimates where replaced with the EB-estimator for that replication, indicated by a superscript † in the cell.

The results in Table 2.3 indicate that, with the saturated model, the Chapman MSE-estimator performs best of the tested estimators, irrespective of the underlying LLM^{S_i}. For $p = 0.01$ it gives a mean value that cannot be rejected to be different from N in 14 out of 15 scenarios. Also, in the scenarios where the Chapman MSE-estimator shows some statistically significant bias for $p = 0.001$ ($S = 5, 7, 10$ and 13), the bias is small in itself and much smaller than in the other estimators. For the IND and 1PD model with large N, the bias of the ML, EB and CFK-estimators is equally large. This unexpected indifference might be due to the modification of elements of n_{abc} that are in the numerator of the ML-estimator, which, as we have seen in Section 2.3.2, is unnecessary. The RL-estimator performs clearly better than the EB- and CFK-estimator, but still shows some statistically significant bias for most scenarios, especially for scenario 13 or when $N = 100$ or 500.

The SEs and RMSEs that correspond to each estimator and scenario in Table 2.3 can be found in Table 2.9 and 2.10 in Appendix 2.6.3.2. These tables show that under an assumed saturated model, the Chapman-estimator not only outperforms the other estimators in terms of bias, but also in terms of SE and RMSE, in particular for smaller N, irrespective of the model specification that was used to generate the contingency tables.

The estimates in Table 2.3 are based on the saturated model, but more restricted models such as those in Eq. (2.18), (2.19) and (2.20) might be assumed. In the next section we will therefore discuss the Chapman MSE-estimator for restricted models.

2.3.3 A generalisation of the Chapman MSE-estimator towards restricted models

The Chapman MSE-estimator for saturated models, as discussed in the previous section, is not necessarily a correct bias-corrected estimator for restricted log-linear models. As an example where the use of the Chapman MSE-estimator for saturated models leads to an incorrect result, consider the 1PD model (2.19) with estimator (2.23). When this estimator uses the modified count vector $\mathbf{n}^{\text{SAT, Chap MSE}} = (n_{111}, n_{110}+1, n_{101}+1, n_{011}+1, n_{100}, n_{010}, n_{001})$ instead of the observed count vector $\mathbf{n}$, this gives

$$\hat{m}_{000}^{\text{Chap MSE',1PD}} = \frac{n_{001}(n_{++0}+1)}{n_{111}+(n_{101}+1)+(n_{011}+1)}.$$

We know that this estimator is not correcting for bias correctly, because the ML-estimator for the 1PD model has the same structure as the LP-estimator, namely the product of two Poisson variables in the nominator (i.e. $n_{001}n_{++0}$) and a single Poisson variable in the denominator (i.e. the sum $n_{111}+n_{101}+n_{011}$). Therefore we should use the same bias-correction as used in the Chapman-estimator, namely

$$\hat{m}_{000}^{\text{Chap MSE, 1PD}} = \frac{n_{001}n_{++0}}{n_{111}+n_{101}+n_{011}+1}. \tag{2.29}$$

This is the correct bias-corrected estimator for the 1PD model. Similarly, for the 2PD model (2.18) with estimator (2.22) we have the bias-corrected estimator

$$\hat{m}_{000}^{\text{Chap MSE, 2PD}} = \frac{n_{001}n_{100}}{n_{101}+1}. \tag{2.30}$$

For the independence model for three sources, a direct solution for the ML-estimator does not exist and therefore we cannot use the approach adopted above for the 1PD and 2PD model as a general solution. Furthermore, for log-linear models with more sources and more source dependencies, the derivations performed by Bishop et al. (1975) become increasingly complex.

Generally, in order to correct for bias, it appears that we should only know which (functions of) observed counts n_{abc} are in the denominator of $\hat{m}_{000}^{\text{ML, LLM}}$, and subsequently adjust these counts to correct for bias. To identify these (functions of) observed counts n_{abc}, we propose to use the Moore-Penrose inverse (Moore, 1920; Penrose, 1955, MPI), that can be used to obtain a "best fit" (i.e. least squares) solution (if any exists) for systems of linear equations.

2.3.3.1 Bias reduction by using the Moore-Penrose inverse

We start with the log-linear model in Eq. (2.5), $\log \mathrm{E}[\mathbf{n}|\mathbf{X}] = \mathbf{X}\boldsymbol{\lambda}$, which is a system of linear equations. A solution for $\boldsymbol{\lambda}$ can be found with the help of the MPI that we write as $\mathbf{Z}^{\text{LLM}} = \left(\left(\mathbf{X}^{\text{LLM}}\right)^{\top}\mathbf{X}^{\text{LLM}}\right)^{-1}\left(\mathbf{X}^{\text{LLM}}\right)^{\top}$, i.e.:

$$\boldsymbol{\lambda}^{\text{MPI}} = \mathbf{Z}^{\text{LLM}} \log\left[\mathbf{n}|\mathbf{X}^{\text{LLM}}\right] = \mathbf{Z}^{\text{LLM}} \log \mathbf{m}. \tag{2.31}$$

For two sources this gives $\mathbf{m} = (m_{11}, m_{10}, m_{01})^{\top}$, $\boldsymbol{\lambda}^{\text{MPI}} = \left(\lambda^{\text{MPI}}, \lambda_a^{A,\text{MPI}}, \lambda_b^{B,\text{MPI}}\right)^{\top}$, $\mathbf{X} = X_{ab} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ and $\mathbf{Z} = \mathbf{Z}_{ab} = \begin{pmatrix} -1 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \end{pmatrix}$. We are interested in a solution for m_{00}, which is $m_{00} = \exp\left\{\lambda^{\text{MPI}}\right\}$, and because λ^{MPI} depends only on the first row of $\mathbf{Z}_{ab}$, only this row is relevant for our purpose. We write this row as the vector $\mathbf{z} = (z_{11}, z_{10}, z_{01})^{\top} = (-1, 1, 1)^{\top}$. Thus (2.31) allows us to write m_{00} as a function of z_{ab} and m_{ab}, i.e.

$$m_{00} = \exp \lambda^{\text{MPI}} = \exp \sum_{ab} z_{ab} \log m_{ab} = \prod_{ab} (m_{ab})^{z_{ab}} = \frac{m_{10} m_{01}}{m_{11}},$$

which corresponds to Eq. (2.2) that led to the LP-estimator, so for two sources $\lambda^{\text{MPI}} = \lambda^{\text{ML}}$. However, for our purpose a more important point is that the first element z_{11} in $\mathbf{z}$ has a negative sign, which indicates that m_{11} is in the denominator of the expression for m_{00}. We have seen in Eq. (2.26), that in order to correct for bias it is important to identify which elements of $\mathbf{n}$ are in the denominator.

This relation between z_{ab} and the LP-estimator also holds for z_{abc} and the SAT, 2PD and 1PD ML-estimators, as defined in Eq. (2.21), (2.22) and (2.23). This can be seen when we specify

$$m_{000}^{\text{LLM, MPI}} = \prod_{abc} (m_{abc})^{z_{abc}^{\text{LLM}}},$$

where z_{abc}^{LLM} depends on the design matrices for restricted log-linear models X_{abc}^{LLM} as defined below Eq. (2.20). The relation between both estimators is further discussed in Frome, Kutner, and Beauchamp (1973), who show that under the Poisson assumption, the maximum likelihood-estimator can be formulated as a properly weighted least

Table 2.4: The value of $\mathbf{z}^{\text{LLM}}$ and $\mathbf{z}^{\text{LLM}}_{<0}$ for each LLM.

Table 2.4a

$\mathbf{m}$	$\mathbf{z}^{\text{SAT}}$	$\mathbf{z}^{\text{2PD}}$	$\mathbf{z}^{\text{1PD}}$	$\mathbf{z}^{\text{IND}}$
m_{111}	1	0	-1/3	-1/2
m_{110}	-1	0	1/3	0
m_{101}	-1	-1	-1/3	0
m_{011}	-1	0	-1/3	0
m_{100}	1	1	1/3	1/2
m_{010}	1	0	1/3	1/2
m_{001}	1	1	1	1/2

Table 2.4b

$\mathbf{m}$	$\mathbf{z}^{\text{SAT}}_{<0}$	$\mathbf{z}^{\text{2PD}}_{<0}$	$\mathbf{z}^{\text{1PD}}_{<0}$	$\mathbf{z}^{\text{IND}}_{<0}$
m_{111}	0	0	-1/3	-1/2
m_{110}	-1	0	0	0
m_{101}	-1	-1	-1/3	0
m_{011}	-1	0	-1/3	0
m_{100}	0	0	0	0
m_{010}	0	0	0	0
m_{001}	0	0	0	0

squares estimator. For the models SAT, 2PD, 1PD and IND, the vector $\mathbf{z}^{\text{LLM}}$ is given in Table 2.4a.

Table 2.4a shows the positive and negative signs in the elements $\mathbf{z}^{\text{LLM}}$ that correspond to the counts n_{abc} in the numerator and denominator in the SAT, 2PD and 1PD ML-estimators in Eq. (2.21), (2.22) and (2.23). It is useful to define $\mathbf{z}^{\text{LLM}}_{<0}$, which is a vector equal to z^{LLM}_{abc} for $z^{\text{LLM}}_{abc} < 0$, and zero otherwise. $\mathbf{z}^{\text{LLM}}_{<0}$ is shown in Table 2.4b for the SAT, 2PD, 1PD and IND model.

For the 2PD and 1PD model the relation between the MPI expression for m_{000} and the ML-estimator is more intricate. For the 1PD model the MPI expression for m_{000} is

$$m^{\text{1PD, MPI}}_{000} = \frac{m_{001}\left(m_{110}m_{100}m_{010}\right)^{\frac{1}{3}}}{\left(m_{111}m_{101}m_{011}\right)^{\frac{1}{3}}}$$

and the ML-estimator in Eq. (2.23) can also be written as

$$\hat{m}^{\text{1PD, ML}}_{000} = \frac{n_{001}\left(m_{110}+m_{100}+m_{010}\right)/3}{\left(m_{111}+m_{101}+m_{011}\right)/3}.$$

The MPI expression for m_{000} is a fraction with geometric means of sets of m_{abc}, both in the numerator and the denominator, while the ML-estimator is a corresponding fraction of arithmetic means of n_{abc}. The same relation can be shown for the SAT and 2PD model. Because a sum of Poisson variables is itself a Poisson variable, and Eq. (2.13) shows that we should add 1 to a Poisson variable in the denominator, where $\mathbf{z}^{\text{LLM}}_{<0}$ provides a distribution of this +1. To illustrate this, in the bias-corrected estimator for the 1PD model in Eq. (2.29), 1 is added to the sum of the three Poisson variables n_{111}, n_{101} and n_{011}. The same result is obtained by subtracting $-1/3$ from each of them, which corresponds to subtracting $\mathbf{z}^{\text{LLM}}_{<0}$ from $\mathbf{n}$. Thus we have a simple formula that can be used to obtain the Chapman-estimator in Eq. (2.10) and the bias-corrected estimators in Eq. (2.21), (2.29) and (2.30), i.e.:

$$\mathbf{n}^{\text{LLM, Chap MSE,}} = \mathbf{n} - \mathbf{z}^{\text{LLM}}_{<0} \tag{2.32}$$

For the IND model Eq. (2.32) implies we should replace n_{111} with $n_{111}+1/2$ to obtain a bias-corrected estimator. We cannot compare this result with a closed form expression of the ML-estimator for the IND model, but an intuitive explanation for this adjustment is that if in the denominator there is a Poisson variable multiplied by a 1/2 as is suggested by the MPI expression, we should add 1 multiplied by a 1/2 to correct for bias as well.

Concluding, in Eq. (2.32) we propose an adjustment that is based on the MPI and can be used for any log-linear model with any number of sources. We have shown for some examples (i.e., for two sources, and for three sources for the models SAT, 2PD and 1PD) that this adjustment works in these instances. The adjustment also works for the saturated model for any number of sources. We provide no proof for other models, such as the model IND for three sources for which no closed form solutions of ML-estimators exist, or restricted models for four or more sources. In the simulation study in the next section we show that also for these models our procedure reduces the bias to a large extent. Finally, we note that the Chapman MSE adjustment of n_{abc} depends on both the log-linear model and the exact inclusion pattern abc, which is more extensive than the information other estimators use.

2.3.3.2 Multiple systems estimation simulation study with restricted models

Table 2.5 shows the results of a simulation study that tests the Chapman MSE-estimators under the different scenarios presented in Table 2.2, and compares them with the other estimators described in Section 2.3.1.1. The estimates are much more accurate because they are based on the same log-linear model that underlies the generation of the contingency table. This is indicated by the LLM^{S_i} in the subscript. Scenarios $10-12$ are removed because they represent scenarios in which the saturated model is the true model, and therefore the results of these scenarios are already provided in Table 2.3.

Table 2.3, 2.10, 2.5 and 2.12 clearly show that, as could be expected, all estimates that are based on the correctly specified model are less biased and have smaller SEs and RMSEs than the estimates based on the saturated model. This difference in bias is caused by the fact that in MSE, finite-sample bias is positive (i.e. m_{000} is overestimated on average) and estimates for models with more parameters have larger variance (Bishop et al., 1975, p. 242). Variance in itself does not lead to biased estimates, but it does inflate bias. With $r = 1,\ldots,R$ and R the number of replications, this inflationary effect can be seen when the bias is written as $(\sum_{r=1}^{R} \hat{m}_{000,r} + n_r)/R - N = (\sum_{r=1}^{R} \exp\left(\hat{\lambda}_r\right))/R - \exp(\lambda)$. When there is some positive bias in the estimate for λ, a larger variance in $\hat{\lambda}$ leads to a further increase of $(\sum_{r=1}^{R} \exp\left(\hat{\lambda}_r\right))/R$ and therefore inflates the bias. This inflation of bias due to increased variance also explains why the Chapman MSE-estimator suffers less from overspecification, which is indicated by the lower root mean squared errors. For example, scenario 7 in Table 2.10 and 2.12 shows that, when instead of the correctly specified two-pair dependence model, the saturated model is assumed, the RMSE of the Chapman MSE-estimator increases

Table 2.5: Simulation study with correctly specified log-linear models, with 100,000 replications, for MSE scenario 1 – 9, 13 – 15 in Table 2.2.

S	N	$\bar{n}$	$\bar{N}^{\text{LLM}^{S_i}, \text{ML}}$	$\bar{N}^{\text{LLM}^{S_i}, \text{EB}}$	$\bar{N}^{\text{LLM}^{S_i}, \text{CFK}}$	$\bar{N}^{\text{LLM}^{S_i}, \text{RL}}$	$\bar{N}^{\text{LLM}^{S_i}, \text{Chap MSE}}$
1	100	79.0	100.5***	100.7***	101.0***	100.7***	99.9***
2	500	332.0	501.5***	501.1***	501.9***	501.3***	499.9
3	10,000	6,587.4	10,001.3***	10,000.8**	10,001.7***	10,001.0***	9,999.7
4	100	77.3	101.2***	101.2***	102.0***	99.9***	100.0
5	500	323.8	503.6***	502.7***	504.3***	499.8*	500.3**
6	10,000	6439.5	10,003.1***	10,002.5***	10,003.7***	10,000.1	10,000.4
7	100	79.1	102.8***†	102.8***	102.9***	100.8***	100.0
8	500	330.5	506.4***	506.2***	505.9***	502.3***	500.3**
9	10,000	6,608.9	10,005.1***	10,005.0***	10,004.8***	10,001.6***	9,999.8
13	1,000	727.0	1,000.7***	1,000.0	1001.2***	1,001.6***	999.5***
14	20,000	10,819.7	20,001.7**	19,997.3***	20,001.9**	20,002.0**	19,996.9***
15	20,000	10,677.5	20,011.6***	20,010.2***	20,008.6***	20,005.2***	20,000.1

$\bar{n}$ gives the mean number of observed units n over all replications. LLM^{S_i} in the superscript indicates that the estimates are obtained under the correctly specified, corresponding log-linear model, as given in the last column of Table 2.2. The superscripts *, ** and *** indicate that we can reject $\hat{N}^{\text{est}} = N$ with a two-sided t-test for p-values = 0.05, 0.01 and 0.001 respectively. A † as superscript indicates that extremely high estimates due to failures were replaced with the corresponding $\hat{N}^{\text{EB},\text{LLM}^{S_i}}$ for that replication.

from 12.3 to 24.9. The other bias-reducing estimators show a much larger increase in RMSE, for example the RMSE of the estimator by Evans and Bonett (1994) increases from 15.4 to 90.0 and the estimator by Rivest and Lévesque (2001) from 13.1 to 44.4.

Particularly interesting is the performance of the Chapman MSE-estimator in the scenarios 1 – 3 and 13 – 14 in the second part of Table 2.5, where the correctly specified independence model is used for estimation. Because the independence model corresponds to the M_t model as defined by Otis et al. (1978), for these scenarios the estimator by Rivest and Lévesque can be directly compared with the Chapman MSE-estimator, while for these scenarios the Chapman MSE-estimator has no justification in a closed-form of the maximum likelihood-estimators. For scenarios 3 and 13 – 14, which have relatively large values for N, both estimators show small but statistically significant bias. However, for scenario 1 and 2, with relatively small N, the Chapman MSE-estimator shows relatively much smaller bias than the other estimators, including the RL-estimator. Finally, in scenario 15 for the correctly specified four-pair dependence model, the Chapman MSE-estimator clearly outperforms the other estimators as well, despite the large N. Together these results are further support for the modification scheme in Eq. (2.32).

Finally, the estimator by Rivest and Lévesque (2001) performs clearly better than the estimators by Evans and Bonett (1994); Cordeiro and McCullagh (1991); Firth (1993) and Kosmidis et al. (2020), but still shows some small but significant bias for

most scenarios, including scenario 1 – 3 and 13 – 14, when using the independence model for estimation.

2.4 Example: Number of homeless people in the Netherlands

A population size estimate of the homeless people in the Netherlands is published annually by Statistics Netherlands. This estimate is an ML estimate that is based on a MSE model that is discussed in detail in Coumans et al. (2017). The estimate is based on a log-linear model that contains three sources and several (categorical) covariates, such as gender (g, 2 categories), age (a, 3 categories), place of living, in- or outside one of the big four Dutch cities (p, 2 categories) and region of origin (o, 3 categories). Together there are 36 subgroups that have observed frequencies denoted as n_{gapo} and an observed frequency with a specific inclusion pattern denoted as $n_{abc,gapo}$. Which sources, covariates and interactions between them are included in the log-linear model, is the result of an Akaike information criterion (Akaike, 1974, AIC) based model selection procedure that is explained in Coumans et al. (2017). Recent work by Silverman (2020) suggests that other model selection approaches based on Bayesian approaches could lead to more robust and stable results, but this is beyond the scope of this paper.

In this practical example, for the years 2009 - 2018, 2020 and 2021, we replicate the model selection and estimation procedure as explained in Coumans et al. (2017). Data for 2019 is unavailable. This gives a series of annual ML estimates for the population size of homeless people in The Netherlands. For each year, the log-linear model that was used to calculate the ML estimate is also used to calculate the corresponding Chapman MSE estimate. This allows us to calculate the difference between the ML and Chapman MSE estimates, all other factors held constant, in a practical example.

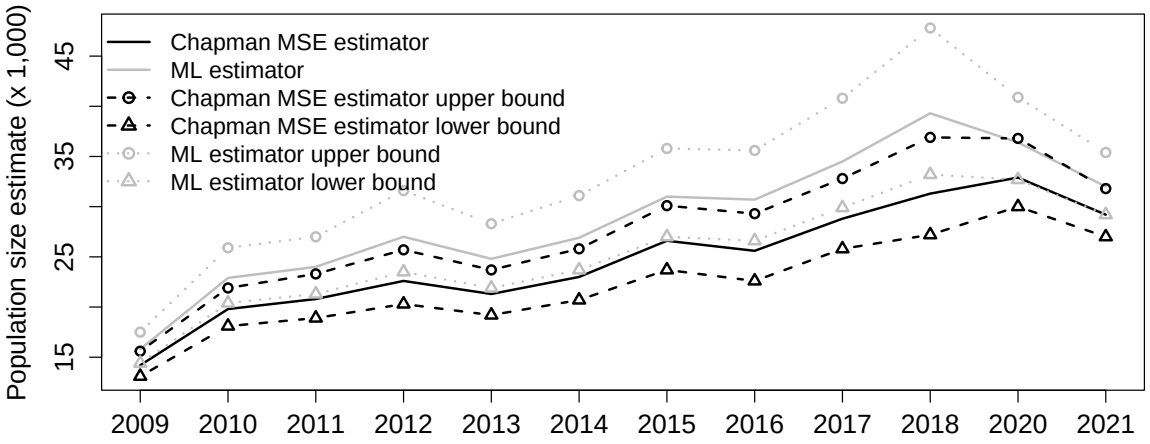
In Figures 2.1a-c below we show, respectively, the original ML estimates and the Chapman MSE estimates of the total number of homeless people, the total number of homeless men and the total number of homeless women, including their two-sided 95% confidence intervals. Note that each figure has its own scale on the y-axis.

Figure 2.1a shows ML and Chapman MSE estimates of the total number of homeless people over time, together with their confidence intervals. The ML estimates are between a minimum of 9.5% and a maximum of 25.5% larger than the Chapman MSE-estimator. The confidence interval of the Chapman MSE-estimators is clearly smaller. Figure 2.1b and 2.1c show that the total annual difference between both estimators, as was observed in Figure 2.1a, is not proportionally divided over men and women. In fact, the Chapman MSE-estimator has, relatively, a much larger impact on the estimate of the number of homeless women, which is the smaller group. For women the difference between the estimates is between a minimum of 19.5% in 2017 and a maximum of 51.2% in 2018.

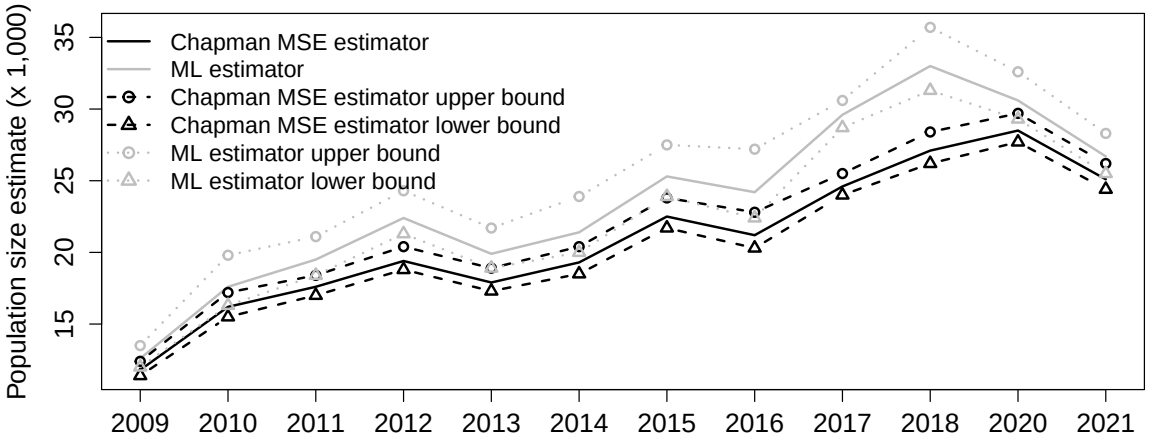
In this practical application the impact of using the Chapman MSE-estimator in-

Figure 2.1: Total number of homeless people, homeless men and homeless women in the Netherlands over the period 2009-2018 and 2020-2021.

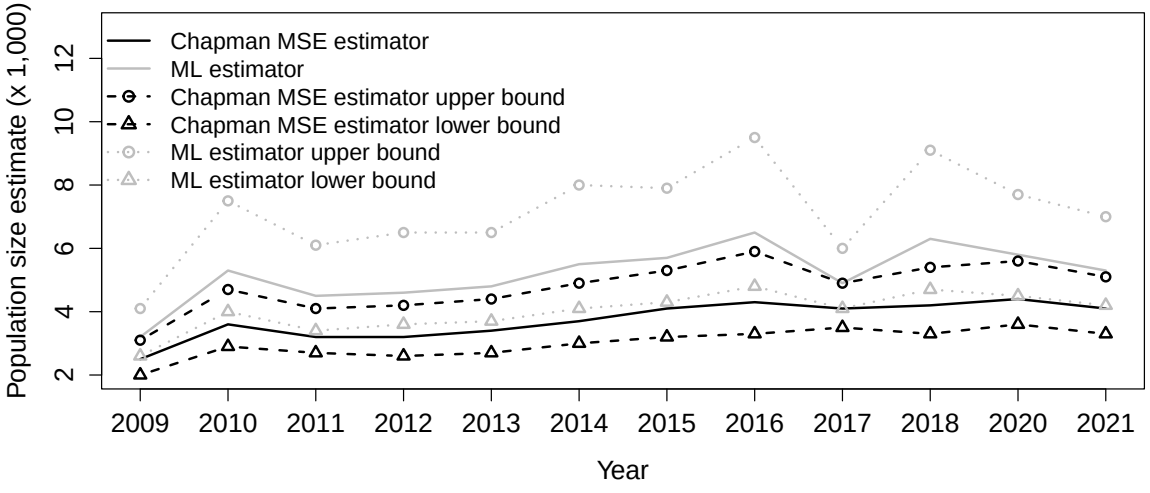
1a: All homeless people



1b: Homeless men



1c: Homeless women



stead of the ML-estimator is larger than the impact we have seen in the simulation studies. The reason for this difference is twofold. First, the scenarios in the simulation studies were set such that the probability of estimation failures was very small, which led to a mean coverage (i.e. $\bar{n}/N$) that was large compared to the coverage in our example of homeless people. Second, the MSE model to estimate the number of homeless people involves the use of (categorical) covariates to control for heterogeneity in inclusion probabilities. Because for some homeless people their background characteristics are missing, the estimation procedure uses an expectation–maximization (EM) algorithm to impute missing data see Coumans et al. (2017) for further details, which for some inclusion patterns may lead to observed frequencies between zero and one. To see why this is important we zoom in on the underlying subgroup estimates for men and women in the year 2021 presented in Table 2.6 below.

Table 2.6 presents 18 subgroups indicated by G_{apo} for both men and women. For each subgroup we show both the total observed count n_{gapo} and the observed count $n_{101,gapo}$ for inclusion pattern 101. This specific inclusion pattern is shown because the selected log-linear model is a 2-pair dependence model, for which Table 2.4 tells us that $n^{\text{Chap}}_{101} = n_{101} + 1$ is the only adjusted observed frequency, while the other elements in n^{Chap}_{abc} are equal to n_{abc}. The difference between $n_{101,gapo}$ and $n^{\text{Chap}}_{101,gapo}$ should therefore explain the difference between N^{ML}_{gapo} and N^{Chap}_{gapo}. This difference is shown in the columns $\Delta_{Mapo} = \hat{N}^{\text{2PD, Chap-3}}_{Mapo} - \hat{N}^{\text{2PD, ML}}_{Mapo}$ and $\Delta_{Wapo} = \hat{N}^{\text{2PD, Chap-3}}_{Wapo} - \hat{N}^{\text{2PD, ML}}_{Wapo}$.

Table 2.6: Estimated number of homeless people in The Netherlands in 2021, separated by men and women and 18 subgroups based on age, living in- or outside one of the four big Dutch cities and country of origin.

	Men					**Women**				
G_{apo}	n_{Mapo}	$n_{101,Mapo}$	$\hat{N}^{\text{2PD, ML}}_{Mapo}$	$\hat{N}^{\text{2PD, Chap-3}}_{Mapo}$	Δ_{Mapo}	n_{Wapo}	$n_{101,Wapo}$	$\hat{N}^{\text{2PD, ML}}_{Wapo}$	$\hat{N}^{\text{2PD, Chap-3}}_{Wapo}$	Δ_{Wapo}
1	1,956	134.07	4,279	4,263	-16	388	8.10	787	678	-109
2	1,283	45.78	4,687	4,464	-223	211	4.03	993	750	-243
3	1,130	37.41	4,458	4,304	-154	164	2.56	760	582	-178
4	516	17.62	1,006	978	-28	97	1.52	170	147	-23
5	496	9.56	2,241	2,065	-176	76	0.90	333	245	-88
6	491	41.02	1,316	1,278	-38	123	3.65	325	264	-61
7	436	36.36	1,072	1,055	-17	102	2.82	243	202	-41
8	350	12.83	1,388	1,302	-86	52	1.11	279	204	-75
9	319	11.04	1,224	1,133	-91	57	1.24	314	226	-88
10	241	7.72	555	533	-22	45	0.66	92	77	-15
11	237	6.07	1,222	989	-233	47	0.63	311	198	-113
12	224	4.84	952	890	-62	35	0.46	142	107	-35
13	201	11.23	685	586	-99	55	1.02	181	130	-51
14	106	2.71	329	274	-55	25	0.29	66	48	-18
15	95	7.82	287	275	-12	28	0.90	89	70	-19
16	91	1.44	561	435	-126	17	0.17	104	65	-39
17	46	1.15	252	194	-58	9	0.14	72	45	-27
18	35	1.90	150	120	-30	11	0.24	50	34	-16
Total	8,253	390.57	26,664	25,138	-1,526	1,542	30.44	5,311	4,072	-1,239

When we compare the columns $\Delta_{Mapo}^{\text{Chap-ML}}$ and $\Delta_{Wapo}^{\text{Chap-ML}}$ in Table 2.6, we see that despite the fact that observed counts of men are larger than those of women, differences in counts of subgroups of men and women are very similar. This can be explained by the smaller observed frequencies for women with inclusion pattern 101, that are sometimes even between zero and one, as can be seen in the columns of $n_{101,Mapo}$ and $n_{101,Wapo}$. Adding 1 to such a small number has a relatively large impact on the population size estimate.

Finally, we note that the Chapman MSE estimates follow a similar trend to the ML-estimates, which is relevant in practice. Only for the period 2018 – 2020 where the ML-estimate is a decrease while the Chapman MSE-estimate is an increase. This might be due to the large ML-estimate in 2018. Furthermore, the estimates and conclusions presented in this section should be treated with some care because for the log-linear model that was used, a regularity condition such as the one for DSE given by Chapman in Eq (2.12) may play a role. The fact that such a regularity condition for MSE is unknown is unfortunate, because some of the subgroups are quite small and so the risk of not meeting a potential regularity condition is not unrealistic. The data on homeless people in The Netherlands that were used for this section is not publicly available due to legal restrictions.

2.5 Discussion

In this paper we have derived the Chapman MSE-estimator and we have shown that, in terms of mean-bias correction, it outperforms a set of other bias-reducing MSE-estimators known in literature. We showed both mathematically and in a simulation study that the mean-bias correction in DSE is best achieved by means of the DSE estimator proposed by Chapman (1951) and later by Rivest and Lévesque (2001). Furthermore we showed how the Chapman-estimator can be derived in a different way than Chapman did. This derivation was extended towards multiple sources, which led to the Chapman MSE-estimator for saturated models. We developed the Chapman MSE-estimator such that it can be applied under both a saturated and restricted model. This generalisation was achieved by using the MPI and for a small set of different restricted models it was proven mathematically that this approach leads to bias-corrected estimators. We used a simulation study to investigate bias in a larger set of restricted models and we found that also for these models the Chapman MSE-estimator shows little or no bias.

The mathematical derivations and simulation studies in this paper show that for any restricted model with three sources or a saturated model with any number of sources, the Chapman MSE-estimator is a bias-corrected estimator. We suspect that this result can be generalised towards any restricted model with any number of sources, although we did not provide a mathematical proof. We think that further research that proves, or disproves, our suspicion would be valuable.

The simulation studies also show that the Chapman MSE-estimator outperforms

other estimators in terms of a smaller size of bias and SE, and thus RMSE, in particular when the estimated log-linear model has more interaction parameters. This advantage is important because in practice the model that is used is usually the result of some model selection procedure, which does not guarantee the selection of the correct model. When such a selection procedure selects a model with irrelevant parameters, this increases the variance of the population size estimate. This increase is less for the Chapman MSE-estimator than for the other estimators considered.

In Section 2.4 we applied the Chapman MSE-estimator to estimate the number of homeless people in The Netherlands for a series of years and compared these estimates with the ML estimates. For each year both estimates are based on the same log-linear model as discussed in Coumans et al. (2017). This comparison showed that the impact of bias-correction can be substantial, e.g., in our example the use of the Chapman MSE-estimator led to a Chapman MSE estimate that was between 9.3% and 25.4% lower for the total number of homeless people in The Netherlands, as compared to the corresponding ML-estimator. This relative difference became even larger, going up to 51%, when we zoomed in on the subgroup of women.

The simulation studies and the example in Section 2.4 show that the difference between the Chapman MSE- and the standard ML-estimator can be substantial. This raises the question whether finite-sample bias correction should not have a more prominent role in the discussion on the robustness of MSE methodology and the accuracy of MSE estimates, which continues up till today (see e.g. Silverman, 2020; Binette & Steorts, 2022).

Finally, a topic that received little or no attention in MSE literature, but what would be valuable to investigate, is regularity conditions. Chapman gave a regularity condition for his DSE estimator, but similar regularity conditions for MSE-estimators are unknown. This topic is also beyond the scope of this paper but we think that this is an important remaining problem for MSE-estimators in general, including the Chapman MSE-estimator.

Software

All simulation studies in this paper are performed in the statistical software program R (R Core Team, 2022). All estimates are obtained with the `glm()` function, with `family = poisson(link = "log")`. Differences between the LP, ML, Chapman, Bailey, EB, RL and Chapman MSE estimates are the sole result of different input vectors $\mathbf{n}^{\text{est}}$. For the IND model the estimation results for the RL-estimator were verified with the function `closedp.bc()` with `m = "Mt"` from the R-package `Rcapture` (Rivest, 2022). The Cordeiro-, Firth- and Kosmidis-estimator ($\hat{N}^{\text{CFK}}$) were also calculated with the `glm()` function, but with the additional settings `method = "brglmFit"` and `type = "correction"`, `type = "AS_mean"` and `type = "MPL_Jeffreys"`, respectively, which are part of the R-package `brglm2` (Kosmidis & Kenne Pagui, 2023). Code for the simulation studies presented in this paper is available at

https://github.com/DaanZult/ChapmanMSE/.

Acknowledgements

The authors thank Jeroen Pannekoek, Peter-Paul de Wolf, Sander Scholtus and Moniek Coumans from Statistics Netherlands for their detailed comments and suggestions on this paper.

2.6 Appendix

2.6.1 Comparison of Taylor approximation and Stephan's inverse factorial approximation

The Taylor expansion that was also used by Bailey (1951) is a widely used approximation approach, but it is not always the most accurate or efficient method to approximate a function. To illustrate that the inverse factorial (IF) expansion (see e.g. Stephan, 1945) used by Chapman (1951) gives more accurate results for $\mathrm{E}\left[\frac{1}{n_{11}}\right]$, given the same number of expansion terms, than a Taylor expansion, we provide a straightforward simulation study. With r replications of $n_{11,r} \sim \text{Poisson}(m_{11})$, we can write the five-term Taylor expansion and IF approximations for $\mathrm{E}\left[\frac{1}{n_{11}}\right]$ as

$$\text{Taylor} \rightarrow \mathrm{E}\left[\frac{1}{n_{11}}\right] = \mathrm{E}\left[\frac{1}{m_{11}} - \frac{(n_{11}-m_{11})}{m^2} + \frac{(n_{11}-m_{11})^2}{m_{11}^3} - \frac{(m_{11}-n_{11})^3}{m_{11}^4} + \frac{(m_{11}-n_{11})^4}{m_{11}^5} - \dots\right]$$

where m_{11} will be estimated by $\hat{m}_{11} = \sum_r n_{11,r}/r$, and

$$\begin{aligned}\text{IF} \rightarrow \mathrm{E}\left[\frac{1}{n_{11}}\right] \approx & \sum_r \left(\frac{1}{n_{11,r}+1}\right)/r + \sum_r \left(\frac{1}{(n_{11,r}+1)(n_{11,r}+2)}\right)/r + \\ & \sum_r \left(\frac{2}{(n_{11,r}+1)(n_{11,r}+2)(n_{11,r}+3)}\right)/r + \\ & \sum_r \left(\frac{6}{(n_{11,r}+1)(n_{11,r}+2)(n_{11,r}+3)(n_{11,r}+4)}\right)/r + \\ & \sum_r \left(\frac{24}{(n_{11,r}+1)(n_{11,r}+2)(n_{11,r}+3)(n_{11,r}+4)(n_{11,r}+5)}\right)/r\end{aligned}$$

Table 2.7 shows the results for both approximation methods and their difference Δ for $m_{11} = 20$ and $r = 1,000,000$. Table 2.7 shows that, for $n_{11,r} \sim \text{Poisson}(m_r = 20)$

Table 2.7: Simulated approximations of $\mathrm{E}\left[\frac{1}{n_{11}}\right]$, with $n_{11,r} \sim \text{Poisson}(m_{11} = 20)$ and r = one million, which gives $\mathrm{E}\left[\frac{1}{n_{11}}\right] \approx \left(\sum_r \frac{1}{n_{11,r}}\right)/r = 0.052805$.

# Terms	Taylor	$\Delta(\mathrm{E}\left[\frac{1}{n_{11}}\right] - \textbf{Taylor})$	IF	$\Delta(\mathrm{E}\left[\frac{1}{n_{11}}\right] - \textbf{IF})$
1	0.050001	0.002804	0.050006	0.002799
2	0.050001	0.002804	0.052507	0.000298
3	0.052505	0.000299	0.052757	0.000048
4	0.052379	0.000426	0.052794	0.000010
5	0.052763	0.000042	0.052802	0.000003

and five or less expansion terms, the IF approximation method used by Chapman (1951) gives a more accurate approximation of $\mathrm{E}\left[\frac{1}{n_{11}}\right] \approx \left(\sum_r \frac{1}{n_{11,r}}\right)/r = 0.052805$ than the Taylor approximation method.

2.6.2 Second-order Taylor approximation of the Lincoln-Petersen-estimator

Here we present an alternative derivation of a bias-reducing LP-estimator. This derivation shows that the Chapman-estimator can be approximated with the well-known Taylor approximation. We write the LP-estimator as a Taylor series approximation. When we start with some function $f(\mathbf{n})$ of the three random variables n_{11},n_{10} and n_{01}, and approximate it around $\mathbf{m}$, this gives:

$$f(\mathbf{n}) = f(\mathbf{m}) + (\mathbf{n}-\mathbf{m})^\top \nabla f(\mathbf{m}) + \frac{1}{2}(\mathbf{n}-\mathbf{m})^\top \nabla\nabla f(\mathbf{m})(\mathbf{n}-\mathbf{m}) + O(\|(\mathbf{n}-\mathbf{m})^\top\|)^2)$$

with

$$\nabla f(\mathbf{m}) = \begin{pmatrix} \frac{\partial f(\mathbf{n})}{\partial n_{11}} \\ \frac{\partial f(\mathbf{n})}{\partial n_{10}} \\ \frac{\partial f(\mathbf{n})}{\partial n_{01}} \end{pmatrix}_{\mathbf{m}}$$

and

$$\nabla\nabla f(\mathbf{m}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{n})}{\partial n_{11}^2} & \frac{\partial^2 f(\mathbf{n})}{\partial n_{11}\partial n_{10}} & \frac{\partial^2 f(\mathbf{n})}{\partial n_{11}\partial n_{01}} \\ \frac{\partial^2 f(\mathbf{n})}{\partial n_{10}\partial n_{11}} & \frac{\partial^2 f(\mathbf{n})}{\partial n_{10}^2} & \frac{\partial^2 f(\mathbf{n})}{\partial n_{10}\partial n_{01}} \\ \frac{\partial^2 f(\mathbf{n})}{\partial n_{01}\partial n_{11}} & \frac{\partial^2 f(\mathbf{n})}{\partial n_{01}\partial n_{10}} & \frac{\partial^2 f(\mathbf{n})}{\partial n_{01}^2} \end{pmatrix}_{\mathbf{m}}$$

Replacing $f(\mathbf{n})$ with $\hat{m}_{00}^{\mathrm{LP}} = \frac{n_{10}n_{01}}{n_{11}}$ gives:

$$\nabla f(\mathbf{n}) = \begin{pmatrix} -\frac{n_{10}n_{01}}{n_{11}^2} \\ \frac{n_{10}}{n_{11}} \\ \frac{n_{01}}{n_{11}} \end{pmatrix}$$

and

$$\nabla\nabla f(\mathbf{n}) = \begin{pmatrix} \frac{2n_{10}n_{01}}{n_{11}^3} & -\frac{n_{01}}{n_{11}^2} & -\frac{n_{10}}{n_{11}^2} \\ -\frac{n_{01}}{n_{11}^2} & 0 & \frac{1}{n_{11}} \\ -\frac{n_{10}}{n_{11}^2} & \frac{1}{n_{11}} & 0 \end{pmatrix}.$$

Therefore, because $\mathrm{E}\left[(\mathbf{n}-\mathbf{m})^\top \nabla f(\mathbf{m})\right] = 0$, we find:

$$\begin{aligned} \mathrm{E}\left[\frac{n_{10}n_{01}}{n_{11}}\right] \approx & \frac{m_{10}m_{01}}{m_{11}} + \\ & \frac{\mathrm{Cov}(n_{10}, n_{01})}{m_{11}} - \frac{m_{10}\mathrm{Cov}(n_{11}, n_{01})}{m_{11}^2} - \frac{m_{01}\mathrm{Cov}(n_{11}, n_{10})}{m_{11}^2} + \\ & \frac{m_{10}m_{01}\mathrm{Var}(n_{11})}{m_{11}^3}. \end{aligned} \tag{2.33}$$

For the Poisson distribution we have $\mathrm{Cov}(n_{ab}, n_{\neq ab}) = 0$ and $\mathrm{Var}(n_{ab}) = m_{ab}$, and for the multinomial distribution we have $\mathrm{Cov}(n_{ab}, n_{\neq ab}) = -Np_{ab}p_{\neq ab}$ and $\mathrm{Var}(n_{ab}) = Np_{ab}(1-p_{ab})$ with $p_{ab} = m_{ab}/N$. Then, for both $n_{ab} \sim \mathrm{Poisson}(m_{ab})$ and the joint set $(n_{11}, n_{10}, n_{01}, n_{00}) \sim \mathrm{Multinomial}(m_{11}, m_{10}, m_{01}, m_{00})$, Eq. (2.33) reduces to:

$$\mathrm{E}\left[\frac{n_{10}n_{01}}{n_{11}}\right] \approx \frac{m_{11}m_{10}m_{01} + m_{10}m_{01}}{m_{11}^2} = \frac{m_{10}m_{01}}{m_{11}}\frac{m_{11}+1}{m_{11}}. \tag{2.34}$$

This implies that $\mathrm{E}\left[\frac{n_{10}n_{01}}{n_{11}}\right]\frac{m_{11}}{m_{11}+1}$ removes the second-order Taylor approximation bias from the LP-estimator, which suggests that multiplying the LP-estimator with $\frac{n_{11}}{n_{11}+1}$, which gives the Chapman-estimator, is an improvement over the LP-estimator.

2.6.3 Tables with SEs and RMSEs

2.6.3.1 DSE

Table 2.8: The SEs and RMSEs for the simulation study presented in Table 2.1.

S	SE^{LP}	SE^{Bailey}	$SE^{EB/CFK}$	$SE^{Chap/RL}$	$RMSE^{LP}$	$RMSE^{Bailey}$	$RMSE^{EB/CFK}$	$RMSE^{Chap/RL}$
1	$27.8^{\dagger}$	20.8	25.8	21.9	$28.3^{\dagger}$	21.2	26.3	21.9
2	28.7	22.2	26.3	23.0	29.3	22.3	26.8	23.0
3	70.2	65.6	68.9	66.5	70.7	65.9	69.3	66.5
4	85.7	78.9	83.3	79.7	86.5	79.0	83.8	79.7
5	460.9	457.9	459.9	458.4	461.2	458.1	460.1	458.4
6	411.3	409.3	410.7	409.6	411.7	409.3	410.9	409.6
7	$109.6^{\dagger}$	45.8	104.2	48.8	$118.9^{\dagger}$	47.5	107.9	49.4

A † as superscript indicates that extremely high estimates due to failures were replaced with the highest Chapman estimate in the simulation sample.

2.6.3.2 MSE with saturated models

Table 2.9: The SEs of the estimates for saturated MSE models, as presented in Table 2.3.

S	N	$SE^{SAT, ML}$	$SE^{SAT, EB}$	$SE^{SAT, CFK}$	$SE^{SAT, RL}$	$SE^{SAT, Chap MSE}$
1	100	61.7	58.8	48.4	38.2	24.1
2	500	104.4	102.9	101.5	94.4	89.9
3	10,000	363.5	363.4	363.3	362.2	361.5
4	100	80.3	77.0	62.7	46.8	26.4
5	500	111.7	109.5	107.3	99.7	94.5
6	10,000	373.7	373.6	373.4	372.2	371.5
7	100	90.6	88.1	65.2	44.2	24.9
8	500	156.5	150.3	143.0	127.9	106.2
9	10,000	391.4	391.2	391.0	389.6	388.6
10	100	83.6	80.8	61.9	43.9	24.1
11	500	138.3	133.4	129.3	112.4	104.0
12	10,000	392.4	392.2	392.0	390.6	389.7
13	1,000	1,280.3	1,270.4	851.1	1,271.7	286.9
14	20,000	4,328.8	4,272.2	4,116.3	4,303.4	3,730.5
15	20,000	7,684.4	7,578.6	7,325.1	7,650.6	4,724.0

A † as superscript indicates that extremely high estimates due to failures were replaced with the highest Chapman MSE estimate in the simulation sample.

Table 2.10: The RMSEs of the estimates for saturated MSE models, as presented in Table 2.3.

S	N	$\text{RMSEs}^{\text{SAT, ML}}$	$\text{RMSEs}^{\text{SAT, EB}}$	$\text{RMSEs}^{\text{SAT, CFK}}$	$\text{RMSEs}^{\text{SAT, RL}}$	$\text{RMSEs}^{\text{SAT, Chap MSE}}$
1	100	63.1	60.1	49.6	38.3	24.1
2	500	106.7	105.3	103.9	94.6	89.9
3	10,000	363.9	363.8	363.7	362.2	361.5
4	100	82.1	78.5	64.0	46.9	26.4
5	500	114.5	112.2	110.0	100.1	94.5
6	10,000	374.1	374.0	373.8	372.3	371.5
7	100	92.9	90.0	66.6	44.4	24.9
8	500	159.8	153.4	146.1	128.3	106.2
9	10,000	391.9	391.8	391.6	389.6	388.6
10	100	85.5	82.5	63.2	44.1	24.1
11	500	141.4	136.6	132.6	112.8	104.0
12	10,000	392.9	392.8	392.7	390.7	389.7
13	1,000	1,296.7	1,285.4	861.6	1,284.9	287.0
14	20,000	4,422.3	4,357.7	4,180.5	4,385.8	3,730.5
15	20,000	7,799.2	7,686.2	7,413.8	7,752.8	4,724.1

A † as superscript indicates that extremely high estimates due to failures were replaced with the highest Chapman MSE estimate in the simulation sample.

2.6.3.3 MSE with restricted models

Table 2.11: The SEs of the estimates for the correct restricted MSE models, as presented in Table 2.5.

S	N	$\text{SE}^{\text{LLM}^{S_i}\text{, ML}}$	$\text{SE}^{\text{LLM}^{S_i}\text{, EB}}$	$\text{SE}^{\text{LLM}^{S_i}\text{, CFK}}$	$\text{SE}^{\text{LLM}^{S_i}\text{, RL}}$	$\text{SE}^{\text{LLM}^{S_i}\text{, Chap MSE}}$
1	100	8.1	8.0	8.1	8.0	7.9
2	500	28.6	28.4	28.6	28.4	28.3
3	10,000	126.1	126.0	126.1	126.1	126.0
4	100	11.6	11.3	11.5	10.8	10.9
5	500	41.4	40.9	41.2	40.2	40.4
6	10,000	164.4	164.3	164.4	164.2	164.2
7	100	15.8	15.1	14.3	13.1	12.3
8	500	48.6	48.2	47.7	46.8	46.0
9	10,000	192.7	192.6	192.5	192.4	192.2
13	1,000	30.1	29.9	30.1	30.1	29.9
14	20,000	255.6	255.4	255.5	255.5	255.4
15	20,000	414.4	414.3	414.1	413.9	413.6

A † as superscript indicates that extremely high estimates due to failures were replaced with the highest Chapman estimate in the simulation sample.

Table 2.12: The RMSEs of the estimates for the correct restricted MSE models, as presented in Table 2.5.

S	N	$\text{RMSE}^{\text{LLM}^{S_i},\text{ ML}}$	$\text{RMSE}^{\text{LLM}^{S_i},\text{ EB}}$	$\text{RMSE}^{\text{LLM}^{S_i},\text{ CFK}}$	$\text{RMSE}^{\text{LLM}^{S_i},\text{ RL}}$	$\text{RMSE}^{\text{LLM}^{S_i},\text{ Chap MSE}}$
1	100	8.1	8.0	8.2	8.1	7.9
2	500	28.6	28.4	28.6	28.5	28.3
3	10,000	126.1	126.0	126.1	126.1	126.0
4	100	11.7	11.4	11.7	10.8	10.9
5	500	41.6	41.0	41.4	40.2	40.4
6	10,000	164.4	164.3	164.4	164.2	164.2
7	100	16.0	15.4	14.6	13.1	12.3
8	500	49.0	48.5	48.1	46.9	46.0
9	10,000	192.7	192.7	192.6	192.4	192.2
13	1,000	30.1	29.9	30.1	30.1	29.9
14	20,000	255.6	255.4	255.5	255.5	255.4
15	20,000	414.6	414.4	414.2	413.9	413.6

A † as superscript indicates that extremely high estimates due to failures were replaced with the highest Chapman estimate in the simulation sample.

Chapter 3

Connecting Correction Methods for Linkage Error in Capture-Recapture

A commonly known problem in population size estimation using registers, is that registers do not necessarily cover the whole population. This may be because they intend to cover part of the population (e.g., students), due to administrative delay or because part of the target population is not registered by default (e.g., illegal persons). One of the methods to estimate the population size in the presence of undercount is the capture-recapture method that combines the information of two or more samples. In the context of census estimation registers are used instead of samples. However, the method assumes that perfect linkage between the registers can be achieved. It is known that this assumption is often violated. In the setting of evaluating the population coverage of a census using a post-enumeration survey, a correction for linkage error was proposed. That correction was later generalized by relaxing some of the newly introduced conditions. However, the new correction method still implicitly assumed that the two registers are of equal size. We introduce a further generalization that includes both previously mentioned correction methods and at the same time deals with registers of different sizes. Specific parameter settings will correspond to the different correction methods. We show that the parameters of each method can be chosen such that the resulting estimates all equal the traditional Petersen estimate (1896) that would theoretically be obtained under truly perfect linkage.

This chapter is published in Journal of Official Statistics: de Wolf, P-P. (PPdW), van der Laan, J. (JvdL) & Zult, D.B. (DZ), 2019. Connecting Correction Methods for Linkage Error in Capture-Recapture. Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 577–597 http://dx.doi.org/10.2478/JOS-2019-0024. Author contributions: PPdW, JvdL and DZ discussed the problem and worked out the idea. PPdW and JvdL did the analysis, PPdW wrote most of the text and JvdL and DZ discussed and edited the text.

3.1 Introduction

The capture-recapture methodology goes back at least to the ecological setting of estimating the size of fish and wildlife populations. The basic idea is to take a first sample (capture), tag or mark the captured animals, return them to their population and take a second sample (recapture). Among the recaptures, some of the animals will be marked, others not. The relation between the tagged and non-tagged animals in the second sample is used to construct an estimate of the total population size (see e.g. Petersen, 1896 and Lincoln, 1930). Since then it was not only used to estimate animal population sizes, but also to estimate undercount in traditional censuses (for an overview see e.g. Fienberg, 1992). More recently, it was used in Gerritse, Bakker, de Wolf, and van der Heijden (2016a) to estimate the undercoverage of registers used for the Dutch Census.

In the original setting, one of the assumptions is that the units can be classified without error to belong to the first sample only, the second sample only or the overlap of the two samples. This assumption was likely to be met, when the marking of the units in the first sample would stick to the animals during the second sampling (no tag-loss). In the setting of estimating the undercount of a register, this assumption is translated to the assumption that units in the two registers can be linked without error, i.e., all links are found and no erroneous links are established. With linking two records, we mean deciding that those records represent the same population unit. Whenever the registers both contain the same reliable unique identifiers like a Social Security Number, it is likely that this assumption holds. However, not all registers contain such a uniform unique identifier. Actually, when considering undercoverage of registers, one can not rely on the existence of such unique identifiers only. Indeed, in order to find units that are not properly registered, one should also use sources that do not have such a unique identifier for all units.

In case a unique identifier is not available, one often relies on probabilistic record linkage techniques like the one developed in Fellegi and Sunter (1969). In this setting the assumption of perfect linkage is not likely to be met in practice. Especially in a large population, two individuals might for example have the same name, leading to a false link, or one individual might be known under two different names, leading to a missed link. This last case would be identical to tag-loss in a classical capture-recapture setting, while the first case can only occur when tags or id-codes can be passed around within the population of interest.

In the presence of linkage errors, the standard capture-recapture estimate of the unknown population size can be biased, (see e.g. Gerritse, Bakker, Zult, & van der Heijden, 2016b). In Ding and Fienberg (1994) the standard capture-recapture estimator is adjusted to correct for linkage errors. In that paper, they considered the situation where a post-enumeration survey (PES) is used to estimate the undercoverage of the population census. See e.g. Wolter (1986) for an explanation of using a PES. Ding and Fienberg assume that the false match that affects the population size estimator most, occurs when a record from the subset of the PES that should not be matched is

actually linked to a record from the subset of the census that should not be matched. In other words, they assume a one-way linkage error, linking PES records to census records. Moreover, they assume that all records in the PES will be linked to a record in the census. Cadwell, Smith, and Baughman (2005) also correct for linkage errors, but they use the concept of 'potential linkage' in a bootstrap procedure to construct a population size interval. Their method is potentially interesting when something like a PES is not available.

Di Consiglio and Tuoto (2015) argue that in the setting of administrative data sources, a one-way linkage direction is not guaranteed. That is, they allow for the possibility that a population unit residing in one administrative data source, but not in the other, to be (incorrectly) linked to a unit in the other administrative data source, irrespective of which data source is called 'the one' and which is called 'the other'. Hence, they propose a two-way correction for linkage error. In their paper, they assume that the probability of a false match is equal in both linkage directions. We will call this the symmetric two-way correction for linkage error. Using the same error probability in both directions, is appropriate in case the two administrative data sources are (approximately) of equal size.

When two registers differ considerably in size, a further extension would be to allow for different error probabilities in the two linkage directions. This would be even more evident when (forced) one-to-one linkage[1] is used. Since the largest source contains units that are not present in the smaller source, in case of one-to-one linkage a subset of those units can never be linked; there are just not enough 'target' records in the smaller source. Records that will never be linked, will also never be linked incorrectly. In other words, a unit in the largest administrative data source has a smaller chance of being falsely linked with a unit in the smaller administrative data source, compared to the other way around. In the current paper we will thus introduce an asymmetric two-way correction for linkage error. The formulation of this asymmetric two-way corrections has three parameters. Choosing specific values for those parameters, the formula can cover the one-way correction and the symmetric two-way correction as well.

The outline of the paper is as follows. We start with explaining the general setting of capture-recapture and probabilistic linkage. In section 3.3 we briefly state the non-corrected estimator, the one-way corrected estimator, the symmetric two-way corrected estimator and an asymmetric two-way corrected estimator. The formula of the asymmetric two-way correction can be viewed as a general estimator in the sense that all introduced estimators can be expressed with this formula. Finally, we unify all estimators by choosing specific 'optimal' parameters. The following section, Section 3.4, shows some simulation results using publicly-available fictitious data on the UK population census. In Section 3.5 we draw conclusions and the appendices contain some technical details.

[1] One-to-one linkage here means that a record from PES is allowed to be linked to one and only one record from the census. Some linkage procedures do not ensure this by default.

3.2 General setting

Let us first introduce the notation that will be used throughout the remainder of this paper. We try to stay close to the notation used in Ding and Fienberg (1994) and Di Consiglio and Tuoto (2015). We also state the general assumptions underlying the capture-recapture methodology when applied with two registers. Note that we will only discuss the situation of two registers that are linked using probabilistic record linkage methods.

3.2.1 Capture-recapture with two registers

Let R_1 and R_2 denote two registers containing units from a common population $\mathcal{X}$ of unknown size $N_{\mathcal{X}}$. Assuming we can identify population units to belong to either one or both of the registers, we get Table 3.1 and Table 3.2. In Table 3.1 the numbers correspond to the unobservable true population counts, whereas the numbers in Table 3.2 are the observed counts *after* the linkage process has taken place and thus depend on the used linkage procedure.

In the tables, the first subscript denotes whether or not a population unit resides in R_1 and the second subscript whether or not a population unit resides in R_2. So, e.g., N_{10} denotes the (unobserved) number of population units that does reside in R_1 but not in R_2. Note that, assuming no duplicates in each R_i, $n_{1+} = N_1$ is the size of R_1, $n_{+1} = N_2$ the size of R_2. Moreover, $N_{\neg i}$ denotes the number of units in the population that do not reside in R_i, i.e., $N_{\neg 1} = N_{\mathcal{X}} - N_1$ and $N_{\neg 2} = N_{\mathcal{X}} - N_2$. Even after the linkage process has taken place, we still cannot observe population units that are included in neither register (i.e., N_{00}). That means that $N_{\neg 1} \geq n_{01}$ and $N_{\neg 2} \geq n_{10}$.

		unit in R_2		
		yes	no	
unit in R_1	yes	N_{11}	N_{10}	N_1
	no	N_{01}	N_{00}	$N_{\neg 1}$
		N_2	$N_{\neg 2}$	$N_{\mathcal{X}}$

Table 3.1: Counts based on population

		unit in R_2		
		yes	no	
unit in R_1	yes	n_{11}	n_{10}	n_{1+}
	no	n_{01}	0	n_{01}
		n_{+1}	n_{10}	n_{++}

Table 3.2: Counts based on linkage process

Using similar notation, we can write the probability that a population unit resides in register R_i as p_i and decompose those probabilities as follows: $p_1 = p_{11} + p_{10}$ and $p_2 = p_{11} + p_{01}$.

The general assumptions in capture-recapture estimation are:

- The population $\mathcal{X}$ is closed, i.e., units can neither enter nor leave the population during the capture-recapture experiment.
- There are no erroneous captures, i.e., only units from $\mathcal{X}$ can be captured.

- There are no duplicates in either register, i.e., units can only be captured once per register.
- The event that a unit resides in R_1 is independent of the event that a unit resides in R_2.
- The probability that a unit resides in R_i is the same for all units in $\mathcal{X}$.
- There is no error in allocating the units to R_1, R_2 or both.

These assumptions imply that $N_{11}/N_1 = N_2/N_{\mathcal{X}}$ or equivalently, $N_{\mathcal{X}} = (N_1 N_2)/N_{11}$. Hence, under perfect conditions a natural estimator would be the one introduced in Petersen (1896): $\hat{N}_{\mathcal{X}} = (n_{1+}n_{+1})/n_{11}$. See Subsection 3.3.1 as well.

3.2.2 Probabilistic record linkage

The probabilistic record linkage technique we will assume in this paper is the one described in Fellegi and Sunter (1969). In their approach, they consider the set of all possible pairs (a, b) of records from R_1 and R_2: $\{(a, b) \mid a \in R_1 \text{ and } b \in R_2\}$. They decompose that set into two disjunct sets. Set $\mathcal{M}$ consisting of all pairs of records of matches and set $\mathcal{U}$ of all pairs of records of non-matches. Hence, e.g., a pair (a, b) in the set $\mathcal{U}$ of non-matches should consist of a record a from register R_1 and a record b from R_2 where a and b refer to two different population units. See Figure 3.2 in Appendix 3.6.1 for a graphical representation of the sets $\mathcal{M}$ and $\mathcal{U}$.

Fellegi and Sunter then describe a model to decide whether an observed pair of records should be allocated to $\mathcal{M}$ or to $\mathcal{U}$. To that end they use so called comparison functions that assign a value to a pair indicating the amount of similarity between the two records. For example, in case of personal data, a comparison function could assign a value zero whenever the name of the person of record a is not exactly equal to the name of the person of record b, and a value of one whenever the names are exactly equal. Obviously, this can be more elaborate, assigning a value between zero and one in case of small spelling mistakes. Different comparison functions can be applied to different variables within a record, which would result in a comparison *vector*.

Selecting a pair of records at random from all possible pairs, the comparison function applied to that selected pair is a random variable. They define the so-called m-probability as the probability that a certain value of the comparison function is found among a pair of records that should belong to the set $\mathcal{M}$ of matches and the u-probability as the probability that a certain value of the comparison function is found among a pair of records that should belong to the set $\mathcal{U}$ of non-matches. Using those probabilities, they assign weights to each possible pair and say that a pair of records is linked whenever the weight is above a some threshold and not-linked whenever that weight is below that threshold. Since this is defined at the level of *pairs* of records, it is possible that several records from register R_1 are said to be linked to the same record in register R_2; whenever a pair has a weight above the threshold, it will be said to be linked. In practice, often a one-to-one linkage is then enforced. One of those

pairs is selected and designated to be a link, while the other pairs are considered to be non-links despite their weight being above the threshold.

In their paper, Fellegi and Sunter consider two error probabilities; the probability of a false link (assigning a pair of records to $\mathcal{M}$ where it should be assigned to $\mathcal{U}$) and the probability of a false non-link (assigning a pair of records to $\mathcal{U}$ where it should be assigned to $\mathcal{M}$). Note that these probabilities are thus defined at the level of *pairs* of records and not on the level of *individual* records. In the description of the correction methods (see Section 3.3) error probabilities are defined at the level of individual records. To be able to discuss the correction methods for linkage error, it is convenient to decompose our registers R_i each into two disjunct sets M_i and U_i. Now M_i consists of all unique records from register R_i that should appear in a pair of matches and U_i of all other unique records from register R_i. Figure 3.2 in Appendix 3.6.1 graphically shows the differences between the sets $\mathcal{M}$, $\mathcal{U}$, M_i and U_i.

Linking registers with many records would lead to a huge number of pairs. Under these circumstances a technique known as blocking is often used to improve efficiency. With blocking, the registers are split into subsets that agree on one or more highly discriminating identifiers and the linkage process is applied within each subset separately. For the sake of simplicity, we will not address the use of blocking in the current paper, since blocking would affect the (estimation of the) m-probabilities in a complex way.

3.3 Estimation of the population size

In this section we will first briefly state the existing estimators for the population size under no linkage error, one-way error correction and symmetric two-way error correction. At the end of this section we will introduce our new asymmetric two-way error correction estimator.

Using the notation from Subsection 3.2.1, we assume that the number of individuals that fall in the four interior cells of Table 3.2 have a multinomial distribution

$$(n_{11}, n_{10}, n_{01}, N_{\mathcal{X}} - n_{++}) \sim \text{Mult}(N_{\mathcal{X}}, p_{11}, p_{10}, p_{01}, p_{00})$$

where $n_{++} = n_{11} + n_{10} + n_{01}$. Like in Ding and Fienberg (1994), we will derive the estimators using the approach of maximizing the conditional likelihood as described in Sanathanan (1972). In that approach the likelihood is written as a product of two likelihoods $L_1(\cdot)$ and $L_2(\cdot)$, where $L_1(\cdot)$ is the likelihood of (n_{11}, n_{10}, n_{01}) for fixed n_{++} and $L_2(\cdot)$ the likelihood of n_{++}, given the cell-probabilities p_{11}, p_{10} and p_{01}. In the conditional approach, first $L_1(\cdot)$ is maximized to derive the maximum likelihood estimates of the cell probabilities, after which the $N_{\mathcal{X}}$ is found that maximizes $L_2(\cdot)$, given the values of p_{11}, p_{10} and p_{01}.

Using that $\mathrm{E}[n_{++}] = \mathrm{E}[n_{1+}] + \mathrm{E}[n_{+1}] - \mathrm{E}[n_{11}] = (p_1 + p_2 - p_{11})N_{\mathcal{X}}$, we derive the

following generic formulation of an estimator of the population total

$$\hat{N}_{\mathcal{X}} = \frac{n_{++}}{\hat{p}_1 + \hat{p}_2 - \hat{p}_{11}} \tag{3.1}$$

In the following subsections we will derive conditional ML estimators of the cell probabilities under different linkage error scenarios.

3.3.1 No linkage error

Under independence and perfect linkage, we would have the following equations for the probabilities of recording population units in the different observed counts n_{ij}

$$p_{11} = p_1 p_2 \tag{3.2}$$
$$p_{10} = p_1 - p_{11} = p_1(1 - p_2) \tag{3.3}$$
$$p_{01} = p_2 - p_{11} = p_2(1 - p_1) \tag{3.4}$$

Using the conditional ML approach we would get the estimators

$$\hat{p}_1 = \frac{n_{11}}{n_{+1}} \qquad \text{and} \qquad \hat{p}_2 = \frac{n_{11}}{n_{1+}}$$

Plugging those estimators into (3.2) and (3.1), the estimator of the population total then becomes after some straightforward calculations

$$\hat{N}^P_{\mathcal{X}} = \frac{n_{1+} n_{+1}}{n_{11}} \tag{3.5}$$

and this is essentially the estimator as described in e.g., Petersen (1896).

3.3.2 One way correction (OC)

In Ding and Fienberg (1994) the situation of linkage error is considered under the assumptions that (using the notation as in Subsection 3.2.2)

(a) A matching pair between records from M_1 and M_2 remains a match with probability $0 < \alpha \leq 1$.

(b) A record from M_1 is matched incorrectly with a record in M_2 with negligible probability.

(c) A false match between records from M_1 and U_2 occurs with negligible probability.

(d) A false match between records from U_1 and M_2 occurs with negligible probability.

(e) Each record from U_1 will be linked with a record in U_2 with common probability $0 \leq \beta < 1$.

The reason for assuming negligible probabilities for (b), (c) and (d) is that in those cases two errors are made; both the correct match is not made and an incorrect match is made. In cases (a) and (e) only one error is made. Note that the probabilities α and β are now defined at *record* level, i.e., different from the probabilities in the Fellegi and Sunter setting (see Subsection 3.2.2).[2] Moreover, note that a large α implies more missed links (in expectation), which in turn leads to an upward bias in the estimator $\hat{N}_{\mathcal{X}}$. A large β implies more false links (in expectation), which would lead to a downward bias in $\hat{N}_{\mathcal{X}}$.

Under the aforementioned assumptions we get the following relations

$$p_{11} = \alpha p_1 p_2 + \beta p_1 (1 - p_2) \tag{3.6}$$

$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta p_1 (1 - p_2) \tag{3.7}$$

$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta p_1 (1 - p_2) \tag{3.8}$$

Note that the 'one-way' correction is reflected in (3.6); the second term on the right hand side only shows the probability of falsely linking (β) a unit that resides in R_1 (p_1) but not in R_2 ($1 - p_2$). The probability of falsely linking a unit that resides in R_2 but not in R_1 is not considered, i.e., only one linkage direction is considered.

The conditional ML estimators are then given by Ding and Fienberg (1994)

$$\hat{p}_1 = \frac{n_{11} - \beta n_{1+}}{(\alpha - \beta) n_{+1}} \qquad \text{and} \qquad \hat{p}_2 = \frac{n_{11} - \beta n_{1+}}{(\alpha - \beta) n_{1+}}$$

Plugging this into (3.6) and (3.1), the population total then can be estimated by

$$\hat{N}_{\mathcal{X}}^{OC} = \frac{(\alpha - \beta) n_{11}}{n_{11} - \beta n_{1+}} \frac{n_{1+} n_{+1}}{n_{11}} = \frac{(\alpha - \beta) n_{11}}{n_{11} - \beta n_{1+}} \hat{N}_{\mathcal{X}}^{P} \tag{3.9}$$

Note that this estimator depends on the parameters α and β which are unknown in practice and should therefore be estimated. This will be discussed in Subsection 3.3.5.

3.3.3 Symmetric two-way correction (SC)

In Di Consiglio and Tuoto (2015) it is proposed to relax the assumption of the one-way correction and to allow a two-way correction. This means that assumption (e) as given in the description of the one-way correction, is relaxed to allow for a unit in U_1 that is not in U_2 still to be (incorrectly) linked to a unit in U_2 as well as to allow for a unit in U_2 that is not present in U_1 still to be (incorrectly) linked to a unit in U_1. Both events occur with the same probability $0 \leq \beta < 1$.

[2] Note that the probabilities in the Fellegi and Sunter setting are sometimes also denoted by α and β. These α and β are thus fundamentally different from the ones used in the current paper.

This results in the following equations

$$p_{11} = \alpha p_1 p_2 + \beta p_1(1-p_2) + \beta p_2(1-p_1) \tag{3.10}$$
$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta p_1(1-p_2) - \beta p_2(1-p_1) \tag{3.11}$$
$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta p_1(1-p_2) - \beta p_2(1-p_1) \tag{3.12}$$

Again, under certain regularity conditions and using the conditional likelihood approach, they derive that the ML estimators are then given by

$$\hat{p}_1 = \frac{n_{11} - \beta(n_{1+} + n_{+1})}{(\alpha - 2\beta)n_{+1}} \qquad \text{and} \qquad \hat{p}_2 = \frac{n_{11} - \beta(n_{1+} + n_{+1})}{(\alpha - 2\beta)n_{1+}}$$

Plugging this into (3.10) and (3.1), the population total then can be estimated by

$$\hat{N}^{SC}_{\mathcal{X}} = \frac{(\alpha - 2\beta)n_{11}}{n_{11} - \beta(n_{1+} + n_{+1})} \frac{n_{1+}n_{+1}}{n_{11}} = \frac{(\alpha - 2\beta)n_{11}}{n_{11} - \beta(n_{1+} + n_{+1})} \hat{N}^{P}_{\mathcal{X}} \tag{3.13}$$

3.3.4 Asymmetric two-way correction (AC)

As a further relaxation of the assumptions, we propose to allow for different probabilities of false links. This means that we allow for a unit present in U_1 but not present in U_2 to be linked to a unit in U_2 with probability $0 \leq \beta_1 < 1$ and a unit present in U_2 but not present in U_1 to be linked to a unit in U_1 but with probability $0 \leq \beta_2 < 1$.

Now the equations for the probabilities of recording population units in the different observed counts become

$$p_{11} = \alpha p_1 p_2 + \beta_1 p_1(1-p_2) + \beta_2 p_2(1-p_1) \tag{3.14}$$
$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta_1 p_1(1-p_2) - \beta_2 p_2(1-p_1) \tag{3.15}$$
$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta_1 p_1(1-p_2) - \beta_2 p_2(1-p_1) \tag{3.16}$$

Under certain regularity conditions, we then get the following ML estimators

$$\hat{p}_1 = \frac{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}}{(\alpha - (\beta_1 + \beta_2))n_{+1}} \qquad \text{and} \qquad \hat{p}_2 = \frac{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}}{(\alpha - (\beta_1 + \beta_2))n_{1+}} \tag{3.17}$$

See Appendix 3.6.2 for a discussion on admissibility to obtain proper values for the probabilities $\hat{p}_1$ and $\hat{p}_2$ in the interval $[0,1]$.

Plugging (3.17) into (3.14) and (3.1), the population total then can be estimated by

$$\hat{N}^{AC}_{\mathcal{X}} = \frac{(\alpha - (\beta_1 + \beta_2))n_{11}}{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}} \frac{n_{1+}n_{+1}}{n_{11}} = \frac{(\alpha - (\beta_1 + \beta_2))n_{11}}{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}} \hat{N}^{P}_{\mathcal{X}} \tag{3.18}$$

Note that this formulation covers all previous situations by choosing appropriate α, β_1 and β_2

- Petersen estimator: $\alpha = 1$ and $\beta_1 = \beta_2 = 0$
- One-way correction: $\alpha = \alpha$, $\beta_1 = \beta$ and $\beta_2 = 0$
- Symmetric two-way correction: $\alpha = \alpha$, $\beta_1 = \beta_2 = \beta$

3.3.5 Linking the correction methods

We consider the Petersen estimator in case of perfect linkage, i.e., knowing the true N_1, N_2 and N_{11}, the 'optimal' estimator and call it the 'true Petersen estimator' (TP)

$$N_{\mathcal{X}}^{TP} = \frac{N_1 N_2}{N_{11}} = \frac{n_{1+} n_{+1}}{N_{11}}$$

Equating the AC estimator to the true Petersen estimator, i.e., setting $\hat{N}_{\mathcal{X}}^{AC} = N_{\mathcal{X}}^{TP}$, we get the following relationship between the parameters

$$\alpha N_{11} + \beta_1 (N_1 - N_{11}) + \beta_2 (N_2 - N_{11}) = \alpha N_{11} + \beta_1 N_{10} + \beta_2 N_{01} = n_{11} \tag{3.19}$$

Note that the left-hand side equals the expected number of links under the model for linkage error.

Let us first explore this relationship under the unrealistic assumption that we know the true N_{11}. A natural choice for the parameter α would then be the fraction of true population matches among the links from the linkage process. We will denote this natural choice by $\check{\alpha}$. Substituting that natural choice in (3.19) and setting $\beta_1 = \beta^{OC}$ and $\beta_2 = 0$, we get

$$\alpha^{OC} = \check{\alpha} = \frac{m_{11}}{N_{11}} \qquad \text{and} \qquad \beta^{OC} = \frac{n_{11} - m_{11}}{N_1 - N_{11}}$$

where m_{11} is the number of true population matches among the links from the linkage process. We will call this choice of parameters the 'optimal OC-parameters'.

In case of the symmetric two-way correction, using the natural choice for α and setting $\beta_1 = \beta_2 = \beta^{SC}$ leads to

$$\alpha^{SC} = \check{\alpha} = \frac{m_{11}}{N_{11}} \qquad \text{and} \qquad \beta^{SC} = \frac{n_{11} - m_{11}}{N_1 + N_2 - 2N_{11}}$$

We will call this choice of parameters the 'optimal SC-parameters'.

In case of the asymmetric two-way correction, we need an additional constraint to uniquely define 'optimal AC-parameters'. In practice, it is convenient to enforce one-to-one linkage in the process. Under that assumption, we can derive the following relationship between the parameters of the asymmetric two-way estimator (see the Appendix 3.6.3 for a derivation)

$$\beta_1 = \frac{(\alpha n_{+1} - n_{11})\beta_2}{(\alpha n_{1+} - n_{11}) - 2\beta_2 (n_{1+} - n_{+1})} \tag{3.20}$$

In case we want to satisfy both (3.20) and (3.19), using the natural $\check{\alpha}$ parameter, we get either

$$\alpha^{AC} = \check{\alpha} = \frac{m_{11}}{N_{11}}, \qquad \beta_1^{AC} = \frac{n_{11} - m_{11}}{2(N_1 - N_{11})} \qquad \text{and} \qquad \beta_2^{AC} = \frac{n_{11} - m_{11}}{2(N_2 - N_{11})}$$

or

$$\tilde{\alpha}^{AC} = \check{\alpha} = \frac{m_{11}}{N_{11}}, \qquad \tilde{\beta}_1^{AC} = \frac{m_{11}N_2 - n_{11}N_{11}}{m_{11}(N_2 - N_1)} \qquad \text{and} \qquad \tilde{\beta}_2^{AC} = \frac{m_{11}N_1 - n_{11}N_{11}}{m_{11}(N_1 - N_2)}$$

where m_{11} again is the number of true population matches among the links from the linkage process. For the second set of parameters ($\tilde{\alpha}^{AC}$, $\tilde{\beta}_1^{AC}$ and $\tilde{\beta}_2^{AC}$) it holds that the $\tilde{\beta}$'s will be undefined in case $N_1 = N_2$. Moreover, when $N_1 \neq N_2$, one of them will be negative, what contradicts the fact that the $\tilde{\beta}$'s should be probabilities. We will hence call the first set of parameters the 'optimal AC-parameters'. Note that, in the case that register R_1 is the largest and hence under one-to-one linkage $N_1 - N_{11} > N_2 - N_{11}$, we get $\beta_1^{AC} < \beta_2^{AC}$ as expected (see discussion in Section 3.1).

According to the error correction model, a false match between a record from U_1 with a record from U_2 occurs with probability β_1 and, independently, a false match between a record from U_2 with a record from U_1 occurs with probability β_2. Considering these events independently, we would count such a link twice. However, enforcing one-to-one linkage, these two events can only happen at the same time. This is reflected in the factor 1/2 in the 'optimal AC-parameters' β_i^{AC}.

Given the true N_{11} and choosing the parameters such that they satisfy equation (3.19) would thus lead to the optimal estimator. Indeed, using the 'optimal OC-parameters', the 'optimal SC-parameters' or the 'optimal AC-parameters' will all yield the same estimator, i.e., the true Petersen estimator TP (with perfect linkage).

Unfortunately, in practice we do not know the true N_{11}. Hence, we need to estimate the α and β_i parameters. As long as the estimated parameters satisfy relation (3.19), the resulting estimates will be exactly the same for all estimators. This would for example be the case when we would estimate the optimal parameters by plugging in some estimate for N_{11}, since N_1, N_2, n_{11} and m_{11} are the same in all settings. Indeed, the resulting estimators would then be given by the simple formula

$$\hat{N}_{\mathcal{X}} = \frac{N_1 N_2}{\hat{N}_{11}} \tag{3.21}$$

where $\hat{N}_{11}$ is a (consistent) estimator of the 'true' overlap between the two registers.

Another possibility would be to use a sample of one of the registers and determine the true matches for that sample. Dividing that number by the sampling fraction would yield a direct estimate of N_{11}. Similarly, we could obtain direct estimates of n_{11} and m_{11}. Note that a direct estimate of n_{11} is needed instead of the original n_{11} to prevent the estimated m_{11} getting larger than the original n_{11}.

Yet another approach would be to use expert knowledge on the linkage errors, e.g., asking experts to give estimates of the parameters. In that case, these expert guesses would not necessarily satisfy relation (3.19) and the estimators could thus yield different values.

3.4 Simulations

Ideally one would like to evaluate estimation methods using 'real life' data. A common way to do so, is to produce out-of-sample predictions. These out-of-sample predictions can be assessed correctly whenever the true values are known at some point. This may for example be the case when predicting a value that can actually be observed 'in the future'. However, in a capture-recapture setting this type of evaluation is not so easy, simply because in general the unobserved population whose size is to be estimated never shows up. Obviously, one could try to collect additional samples and count the number of new records that show up, but one can never be sure that no individual was left unobserved. Therefore, a more appropriate way to evaluate the effectiveness of population size estimators in the capture-recapture setting is by means of a simulation study.

Simulation studies of course have the disadvantage that they use artificial data. This disadvantage can somewhat be reduced by using a (subset of) properly privacy protected real data as an "artificial" population. That way, the data will contain real linkage keys, real covariates and will have realistic measurement errors, while the true population size is known.

Simulated data has the advantage that the population as well as the registers are completely known; we know all the entries of Table 3.1 as well as Table 3.2. We can thus easily determine how well the estimators approximate the true population size. An additional advantage is that we can derive the 'optimal' Petersen estimator; the Petersen estimator with truly no linkage error. Since this is the maximum likelihood estimator using population information, the resulting estimate is the best one could get. We will call this estimator the True Petersen estimator and use it as a benchmark for our other estimators in our simulations. The True Petersen estimator is thus based on the counts in Table 3.1 and does not equal the Petersen estimator one would get in practice using the counts from Table 3.2.

Since ensuring that the parameter estimates satisfy relation (3.19) will result in the same estimates of the population size for all estimators introduced in section 3.3, we will concentrate on different ways to estimate the parameters. We will use different methods to estimate N_{11} and m_{11} and plug those estimates into the formulas of our 'optimal' parameters, to show empirically that these estimates indeed lead to the same estimate of the population total.

3.4.1 Setup

For the simulation we will make use of the fictitious data based on the UK population census as created for the ESSnet DI (McLeod, Heasman, & Forbes, 2011). The ESSnet DI was a European project on data integration (Record Linkage, Statistical Matching, Micro Integration Processing), running from 2009 to 2011. We used three files from that dataset; the files `Person` (a fictional list of persons, acting as the population), `CIS` (fictional observations from a Customer Information System, being a combination of

tax and benefit data) and PRD (fictional observations from the Patient Register Data of the National Health Service). The Person dataset comprised of 26,625 individuals, the CIS has a coverage probability of that population of $\tau_1 = 0.930$ and the PRD of $\tau_2 = 0.924$.

To reduce computation time and to be able to apply the linkage process without blocking, we repeatedly constructed a smaller population and corresponding registers from those files, using the following steps

1. Draw a simple random sample without replacement of size 10,000 from Person. This will be our population $\mathcal{X}$ with size $N_{\mathcal{X}}$.

2. Select the records from CIS that are present in population $\mathcal{X}$ to get register R_1.

3. Select the records from PRD that are present in population $\mathcal{X}$. Randomly select a fraction f of those records to get register R_2.

This way we obtained multiple instances of a population and the corresponding registers where one of them covers the population for about 93% and the other for about f times 92.4%. Note that, for small values of f, the two registers differ substantially in size.

In Di Consiglio and Tuoto (2015) several linkage scenarios were mentioned; a bronze, a silver and a gold scenario. In the current paper we will only use their silver scenario, i.e., we only use the full date of birth (day (DB_D), month (DB_M) and year (DB_Y)) as key variables in the linkage process. We have chosen the silver scenario because it allows for two types of linkage error. Firstly, two different individuals may have the same date of birth and therefore may be falsely linked. Secondly, due to some measurement errors, an individual that is in both samples may be falsely not linked. Names and surnames would have been better discriminating identifiers, but in the absence of those variables (e.g., due to privacy restrictions), the full date of birth is still reasonably discriminating.

For the comparison function of the probabilistic record linkage process (see Subsection 3.2.2), we simply used 'equality' on all key variables separately. That is, whenever two records a and b are compared, the comparison function for key variable V_i is 1 when $V_i(a) = V_i(b)$ or 0 when $V_i(a) \neq V_i(b)$. Whenever V_i is missing in at least one of the two records, the comparison function is defined to be 0 as well. To perform the probabilistic record linkage as described in Subsection 3.2.2, we used our own R-code. In that code we also forced one-to-one linkage. See https://github.com/djvanderlaan/reclin for the R-package reclin that we used.

We implemented four methods to obtain values for the N_{11}, m_{11} and n_{11} needed in the formulas for the 'optimal' parameters

A Since we use simulated data, we know the true m_{11} and N_{11} by design. The n_{11} follows from the linkage process.

B Using the EM-algorithm, (see e.g. Herzog, Scheuren, & Winkler, 2007) on the complete registers to estimate the posterior m-probabilities. Those posterior

probabilities were used to estimate the m_{11} and N_{11}. The n_{11} follows from the linkage process.

C Using a sample of size 200 from the smallest register of which we determine the true match-status. Using that sample we fitted a logistic model (see the Appendix 3.6.4 for more information on the used model) and used that to predict the m-probabilities for the complete registers. Those posterior probabilities were used to estimate the m_{11} and N_{11}. The n_{11} follows from the original linkage process.

D Using a sample of size 200 from the smallest register of which we determine the true match-status. Using that sample we calculated the direct estimates of n_{11}, N_{11} and m_{11} for the complete registers.

In methods B and C, summing the posterior m-probabilities over all linked pairs yields an estimate of m_{11}, whereas summing those probabilities over all possible pairs yields an estimate of N_{11}. For a definition of posterior m-probabilities and why summing them is appropriate, we refer to Fellegi and Sunter (1969). Methods B, C and D serve to illustrate how information available during an actual record linkage process can be used to correct the estimator for linkage errors. As long as the sample used in methods C and D is a representative sample of possible record pairs, these methods should give unbiased estimates of N_{11}, m_{11} and n_{11}. Other methods or refinements of these methods that might give more precise estimates are possible. However, finding such refinements is not the main focus of this paper; we want to show that even relatively simple methods can already correct for bias due to linkage errors.

With those estimated sizes, we then used the formulas for the 'optimal' parameters as derived in section 3.3.5 to get estimates of the population size. As discussed in that section, we expect to obtain exactly the same estimates for all approaches (OC, SC and AC).

3.4.2 Results

For three different values of f, we performed 100 replications of the procedure mentioned in the previous subsection and, as expected, we indeed found that all 'optimal' parameters led to the same estimates in all four methods. In Table 3.3 the mean, median and standard deviation over the 100 replications is given for the difference between the estimates of the population size and the actual population size $N_{\mathcal{X}} = 10,000$, for the estimators TP (method A, the benchmark), P (Petersen, using the counts from the linkage process), EM (method B), model (method C) and sample (method D) . Note that TP and P are both based on Petersen's formula (Petersen, 1896), but TP is using the (in practice unobservable) true population counts, whereas P uses the observed counts.

The first thing to notice, is that the Petersen estimator using the observed counts indeed leads to a heavily biased estimate of the population size, due to the linkage

f		TP $- N_{\mathcal{X}}$	P $- N_{\mathcal{X}}$	EM $- N_{\mathcal{X}}$	model $- N_{\mathcal{X}}$	sample $- N_{\mathcal{X}}$
0.15	mean	10.9	1,033.9	-1,874.2	-13.7	10.9
	median	1.3	1,032.3	-1,767.4	-5.2	-18.2
	st. dev.	68.0	124.7	1,211.5	283.2	212.9
0.50	mean	6.0	1,186.2	-1,826.8	12.3	11.6
	median	4.5	1,186.8	-1,859.0	-30.1	29.4
	st. dev.	32.6	58.4	794.2	263.9	202.2
0.90	mean	5.6	1,398.0	-1,805.8	19.4	53.8
	median	6.7	1,397.5	-1,851.4	9.8	15.7
	st. dev.	11.8	39.3	791.1	302.4	205.8

Table 3.3: Mean, median and variance of the difference with $N_{\mathcal{X}} = 10,000$ of each estimator over the 100 replications, for different relative sizes f of the second register, and sample size 200.

errors that are present. Moreover, we see that the EM-based estimator (method B) has a very large variance compared to the other estimators and at the same time has a larger bias. This indicates that this method is not well suited to be used for correcting linkage error.

Varying the relative size of the second register (i.e., the f) does not really influence the correction for linkage error. Indeed, the bias as well as the variance of those estimators seems to be more or less the same in all situations.

In case the registers include a unique identifier for some of the records, the identifier could be used as an alternative for taking a sample, under the assumption that the absence of the identifier is not (too) selective. When such a unique identifier is not present, it could in practice be quite costly to determine the true match status of pairs. Hence, probably only a small sample would be considered by a National Statistical Institute and that's why we used a relatively small sample from the second register for methods C and D.

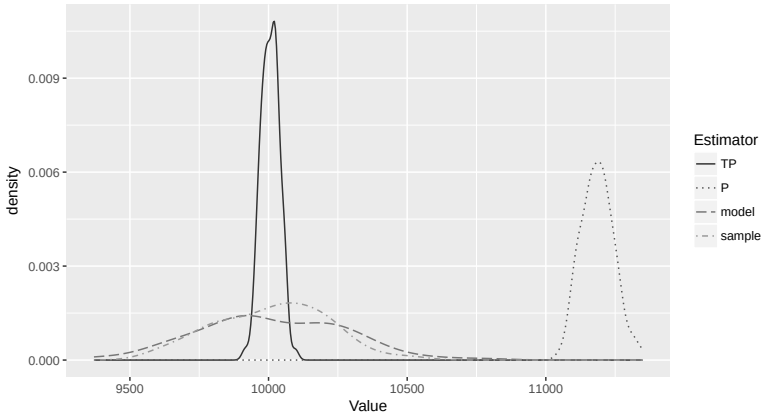
Figure 3.1 shows a smooth estimate of the distribution of the estimators for $f = 0.5$. For the other values of f the distributions look similar. We did not plot the EM-based estimator in this figure to be able to see more clearly the differences between the other estimation methods.

The figure again shows clearly that the Petersen estimator using the counts from the linkage process has a large bias (due to linkage error) and that the model and sample estimators nicely correct for that. The TP estimates are obviously performing the best, since they use the true knowledge about the number of matches. However, in practice that estimator is not possible.

3.5 Conclusions

In estimating the population size using capture-recapture, linkage errors (false links and missed links) affect the Petersen estimator. Indeed, the Petersen estimator then

Figure 3.1: Distributions of the P, TP, model based and sample based estimators for $N_{\mathcal{X}} = 10,000$, $f = 0.5$ and sample size 200.



becomes heavily biased. To reduce that bias, some correction methods have been proposed in the literature. These methods introduce some additional parameters that should reflect the probability of occurrence of the two possible types of linkage error. They then model how linkage errors occur and use those error-probabilities to incorporate that model into the estimator. In this paper we have introduced a general formulation for such a correction method. That general formulation incorporates all previously introduced correction methods of that type as special cases.

Looking more closely to the general correction method, it turned out that the parameters could actually be chosen in such a way that the general estimator equals the optimal estimator; the Petersen estimator with known number of true matches. These 'optimal' parameters can be estimated using different methods. We have shown that for at least two methods, the results improve the traditional Petersen estimator considerably. Those two methods make use of a relatively small sample for which the true match status of the records needs to be determined. More refined methods might even improve more and lead to estimators with smaller variances.

We have shown that it is possible to chose 'optimal' parameters, such that all adjustment methods lead to exactly the same estimates. This reduces the need for making a choice on the error linkage model. However, in case the probabilities are estimated in a different way (e.g., by means of expert opinions), the different linkage error models will lead to different estimates. We have not investigated this further in the current paper.

In case it is not possible to make use of a sample to estimate the 'optimal' parameters, the general correction method could still be useful. In that situation, the model for the occurrence of the linkage errors should be assessed to estimate the error probabilities. We would like to note that the model assumes that 'double errors' occur with negligible probabilities. With 'double errors' we mean errors like missing a true match of a record and at the same time linking that record incorrectly to some record in the other register. In estimating the error-probabilities this should be taken

into account in some way, because in practice such double errors do occur and would influence the error probabilities.

Using covariates or linking more than two registers would lead to more elaborate methods to estimate the population size in the presence of undercoverage. In these cases, more complex loglinear or Poisson models can be used to obtain a capture-recapture estimate. Similarly, the Fellegi and Sunter based linkage procedure can also be applied more elaborately, e.g., by making use of blocking(s). This would affect the (estimates of the) posterior m-probabilities. In our view, the ideas expressed in the current paper, as well as the introduced general formulation of the linkage error correction methods, will lead to a better understanding of the implications of such extensions and will be of help in deriving new, linkage error correcting, consistent estimators of the population size.

3.6 Appendix

3.6.1 Sets defined in the setting of probabilistic record linkage

Let R_1 be a register with records numbered $\{1,2,3,\ldots,10\}$ and R_2 a register with records numbered $\{1,2,3,\ldots,15\}$. The total number of *pairs* (a,b) that can be constructed from the *records* of those registers is $10 \times 15 = 150$. Figure 3.2 shows all possible pairs. Moreover, an example of the set $\mathcal{M}$ of pairs of matching records and the set $\mathcal{U}$ of pairs of non-matching records is shown in that figure. In the example, the number of pairs in $\mathcal{M}$ is 8 and the number of pairs in $\mathcal{U}$ is 142.

We can write each register as the union of two disjoint sets, $R_i = M_i \cup U_i$, where the disjoint sets of unique records are given by

$$M_1 = \{1,3,4,5,6,7,8,9\} \qquad U_1 = \{2,10\}$$
$$M_2 = \{2,3,4,6,8,9,10,13\} \qquad U_2 = \{1,5,7,11,12,14,15\}$$

3.6.2 Admissibility of asymmetric two-way correction estimators $\hat{p}_i$

The estimators for the probabilities p_i in case of the asymmetric two-way correction approach should obviously be within $[0,1]$. This puts some restrictions on the parameters α, β_1 and β_2.

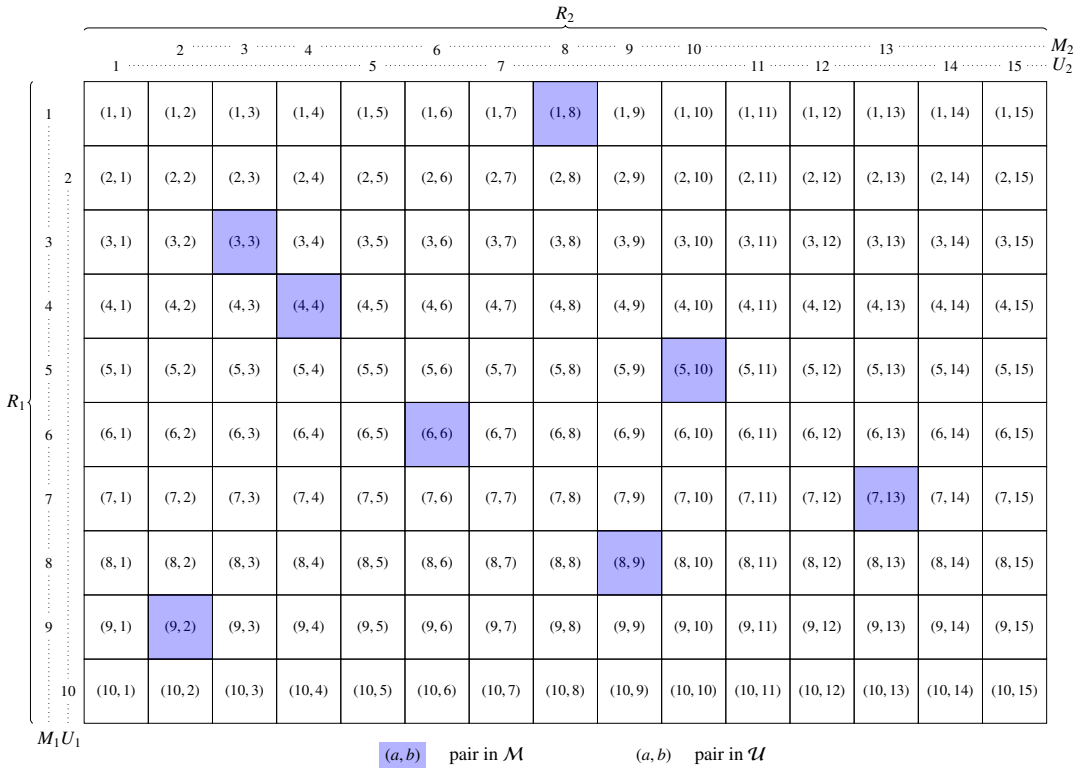
To ensure that the estimators are non-negative, straightforward calculations lead to the condition that either

$$\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \qquad \text{and} \qquad \beta_1 + \beta_2 < \alpha \tag{3.22}$$

or

$$\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \qquad \text{and} \qquad \beta_1 + \beta_2 > \alpha \tag{3.23}$$

Figure 3.2: Graphical representation of the sets of pairs defined in subsection 3.2.2



Additionally, ensuring that both probabilities are not larger than one, leads under (3.22) to the condition

$$\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \wedge n_{+1}) \tag{3.24}$$

and under (3.23) to the condition

$$\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \vee n_{+1}) \tag{3.25}$$

where $n_{1+} \vee n_{+1}$ equals the maximum of n_{1+} and n_{+1} and $n_{1+} \wedge n_{+1}$ the minimum of n_{1+} and n_{+1}.

Summarizing, we need either

$$\left.\begin{aligned} &\beta_1 + \beta_2 < \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \wedge n_{+1}) \end{aligned}\right\} \tag{3.26}$$

or

$$\left.\begin{aligned} &\beta_1 + \beta_2 > \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \vee n_{+1}) \end{aligned}\right\} \tag{3.27}$$

Assuming R_1 to be the largest data set, i.e., $n_{1+} > n_{+1}$, the set of conditions (3.26) is equivalent to

$$\left.\begin{array}{l} \beta_1 \geq (n_{11} - \alpha n_{+1}) \Big/ (n_{1+} - n_{+1}) \\ \beta_1 + \beta_2 < \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \end{array}\right\} \tag{3.26'}$$

and the set of conditions (3.27) to

$$\left.\begin{array}{l} \beta_2 \geq (\alpha n_{1+} - n_{11}) \Big/ (n_{1+} - n_{+1}) \\ \beta_1 + \beta_2 > \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \end{array}\right\} \tag{3.27'}$$

Assuming the two data sets to be of equal size, i.e., $n_{1+} = n_{+1}$, the set of conditions (3.26) is equivalent to

$$\left.\begin{array}{l} \alpha \geq n_{11} \Big/ n_{1+} \\ \beta_1 + \beta_2 < \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \end{array}\right\} \tag{3.26''}$$

and the set of conditions (3.27) to

$$\left.\begin{array}{l} \alpha \leq n_{11} \Big/ n_{1+} \\ \beta_1 + \beta_2 > \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \end{array}\right\} \tag{3.27''}$$

3.6.3 Enforcing one-to-one linkage

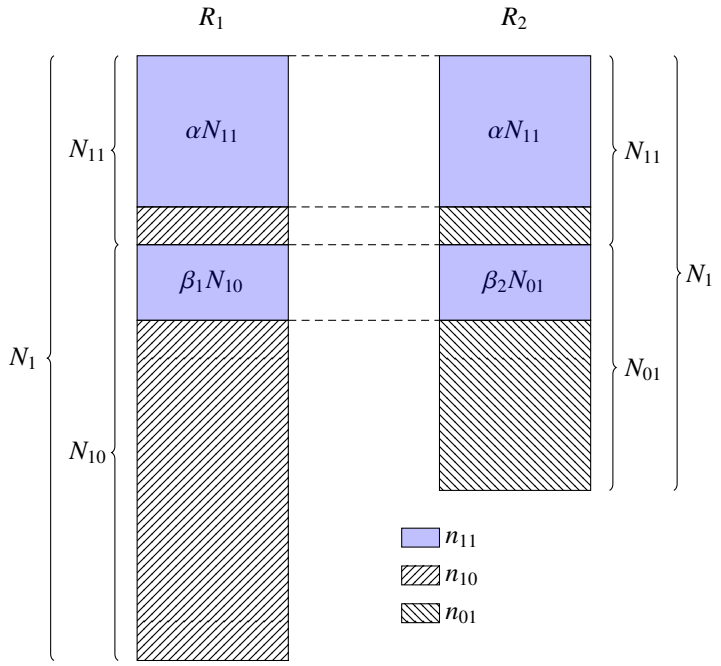
In our asymmetric two-way correction method, we have three parameters; α, β_1 and β_2. In case we enforce one-to-one linkage, we can actually do with two, because in that situation we can write β_1 as a function of α and β_2.

In Figure 3.3 the relation between (expected) counts based on the population and based on linkage are shown in the situation where we potentially would like to apply the asymmetric two-way correction with enforced one-to-one linkage. Under the assumption of one-to-one linkage, it should hold that $\beta_1 N_{10} = \beta_2 N_{01}$, as can be seen in the figure. Noting that $\mathrm{E}[N_{10}] = p_1(1-p_2)N_{\mathcal{X}}$ and $\mathrm{E}[N_{01}] = p_2(1-p_1)N_{\mathcal{X}}$ and plugging in the estimators $\hat{p}_1$ and $\hat{p}_2$ from (3.17), we can derive the following relation

$$\beta_1 = \frac{(\alpha n_{+1} - n_{11})\beta_2}{(\alpha n_{1+} - n_{11}) - 2\beta_2(n_{1+} - n_{+1})} \tag{3.28}$$

Note that, assuming equal sizes of the two registers, i.e., $n_{1+} = n_{+1}$, equation (3.28) yields $\beta_1 = \beta_2$. That is, we would obtain the situation in which the symmetric two-way correction is applicable.

Figure 3.3: Relations between counts based on population and based on one-to-one linkage



Moreover, from (3.28) it follows that

$$\alpha > 2\beta_2 \text{ and } n_{1+} > n_{+1} \implies \beta_1 < \beta_2$$
$$\alpha > 2\beta_2 \text{ and } n_{1+} < n_{+1} \implies \beta_1 > \beta_2$$

as expected (see discussion in Section 3.1).

3.6.4 Estimation of the matching probabilities using logistic regression

For a sample of records from the smallest register it is assumed that the true match status can be determined, i.e., we assume that it is known whether or not the record should be linked to a record from the larger register and if so with which record it should be linked. Therefore, for a subset of all pairs generated in the linkage process, the true match status is known. The goal of the logistic regression model is to predict the probability that this pair is a true match, based on properties of the record pair.

In the regression model the following covariates are used

1. The result of the comparison of the linkage variables. In this case the linkage variables are the three elements of the date of birth; day (`DB_D`), month (`DB_M`) and year (`DB_Y`). These variables are binary; both records of the pair agree on the variables (true) or not (false). If in at least one of the records a variable is missing, we consider it a disagreement (false).

2. Whether or not the pair is selected when enforcing one-to-one linkage (LNK). This is also a binary variable which is false when there is a more likely match for one or both of the records. This variable is a strong predictor for true matches.

The target variable is the true match status (a binary variable). All variables are added as main effects. No interactions are used in the model. The model is estimated using the sampled pairs and then used to calculate predictions of the matching probability for all pairs.

To estimate m_{11} the probability that a pair is a true match given that a pair has been linked is needed, and to estimate N_{11} the probability of a true match given that a pair has been linked or has not been linked is needed. Therefore, as long as the sample is representative for the set of pairs, using only LNK should be enough to obtain unbiased estimates of N_{11} and m_{11} (n_{11} follows directly from the linkage procedure). Adding additional variables to the regression model, such as DB_D, DB_M and DB_Y in this case, could lead to a reduction of the variance of the estimators when the probability of a false link depends on this variable. However, as this is strongly data set dependent and as the main goal of the correction method is the removal of the bias, additional covariate candidates were not investigated.

Chapter 4

A General Framework for Multiple-Recapture Estimation that Incorporates Linkage Error Correction

The size of a partly observed population is often estimated with the capture-recapture model. An important assumption of this model is that sources can be perfectly linked. This assumption is of relevance if the identification of records is not obtained by some perfect identifier (such as an id code) but by indirect identifiers (such as name and address). In that case, the perfect linkage assumption is often violated, which in general leads to biased population size estimates. Initial suggestions to solve this use record linkage probabilities to correct the capture-recapture model. In this article we provide a general framework, based on the standard log-linear modelling approach, that generalises this work towards the inclusion of additional sources and covariates. We show that the method performs well in a simulation study.

This chapter is published in the Journal of Official Statistics: Zult, D.B. (DZ), de Wolf, P-P. (PPdW), Bakker, B.F.M. (BB) and van der Heijden, P.G.M. (PvdH), 2021. A General Framework for Multiple-Recapture Estimation that Incorporates Linkage Error Correction. Journal of Official Statistics, Vol. 37, No. 3, 2021, pp. 699–718 https://doi.org/10.2478/jos-2021-0031. Author contributions: BB proposed the problem, DZ and PPdW discussed the problem and worked out the idea. DZ did the analyses and wrote most of the text and BB and PvdH discussed and edited the text.

4.1 Introduction

Capture – recapture (CR) estimation provides a standard approach to estimate the size of a population, including the unobserved part (Petersen, 1896; Fienberg, 1972; Bishop et al., 1975). These models are also known under other names, such as dual - system, multiple system and mark – recapture estimation models (see e.g. International Working Group for Disease Monitoring and Forecasting, 1995a). Dual-system (DS) estimation uses two sources and multiple system (MS) estimation uses three or more sources (see e.g. Fienberg, 1972). A source refers to a set, list or register of records. We assume that each record represents a unit that is unique to that source and belongs to the target population. When the combination of available sources does not cover the full target population, under specific assumptions as described in Wolter (1986), CR models can be used to estimate the size of the missing part of the population. One of the assumptions in CR models is that records can be perfectly identified over sources as belonging to the same unit or not. This allows an accurate linkage of records and sources into one combined source. If a perfect identification of units is not possible there is a non - zero probability that records will be falsely linked (a mismatch), or falsely not linked (a missed match) and the resulting population size estimate (PSE) is generally biased (see e.g. Wolter, 1986; Chao, 2001; Z. Chen & Kuo, 2001; Cadwell et al., 2005; Bakker et al., 2017). A first step of a solution to this problem was provided by Ding and Fienberg (1994, D&F). For the linkage of two sources S^1 and S^2 they define five different linkage error types:

(1) A missed link between the same unit that is in both S^1 and S^2.

(2a) A false link between two different units that are both in S^1 and S^2.

(2b) A false link between a unit that is in S^1 and S^2 and a different unit that is only in S^2.

(2c) A false link between a unit that is in S^1 and S^2 and a different unit that is only in S^1, and

(2d) A false link between two different units that are in S^1 and S^2.

Linkage error type (1) concerns a missed match while type (2a) – (2d) concern different types of mismatches. To simplify the model, D&F assume that linkage error types (2a) – (2c) are negligible because they require a double linkage error. Therefore, they derive a model that corrects for the two remaining linkage error types (1) and (2d). The D&F model requires a rematch study. This is a study that checks whether a subset of record linkages and non – linkages is correct or not and is usually carried out by a clerical review. This subset is assumed to be representative for the entire population. The D&F model uses the rematch study to obtain different sorts of linkage error probabilities. Note that linkage errors refer to record linkage errors that occur during source linkage. A record linkage is the linkage between two records in two sources.

Source linkage refers to the linkage of records in two or more sources. The D&F model is extended by Di Consiglio and Tuoto (2015, DC&T_15) and Di Consiglio and Tuoto (2018, DC&T_18). They showed that D&F only explicitly consider the probability of a record in S^1 to be falsely linked to a record in S^2, while a record in S^2 can just as well be falsely linked to a record in S^1. Therefore, DC&T_15 derive a model that takes both options into account. Further progress is presented by de Wolf et al. (2019, WLZ), who showed that both D&F and DC&T_15 implicitly assume that S^1 and S^2 are of equal size. This is important, because the probability of a false link increases or decreases when the number of potential record linkages increases or decreases, which depends on the size of both sources. Therefore, WLZ derive a model that takes these different source sizes into account. This progress in DC&T_15 and WLZ is restricted to linkage error type (1) and (2d). WLZ take one more step and derive a (what we refer to as D&F+) model that takes all five linkage error types into account. We will see in Section 4.3 that this D&F+ model is also less complex and can be used to generalize the model even further. Despite the progress the D&F+ model still suffers from two major shortcomings:

(1) It is unclear how to perform statistical inference with respect to covariates in the model.

(2) The D&F+ model is only defined for two sources and not for three or more.

These two shortcomings are important in case captures are covariate and/or source dependent, because it implies that linkage error correction is not possible, if covariates and/or additional sources are also required to correct for covariate and/or source dependencies. The linkage errors also cannot be modelled explicitly in case of recapture-prone or recapture-adverse populations (see e.g. Chatterjee & Mukherjee, 2018). If there are two sources these linkage error probabilities might still be incorporated in the derivation of a closed form maximum likelihood estimator when the sources are independent. However, this derivation becomes increasingly complicated when covariates and additional sources are added, and it is unclear how to do statistical inference in this situation. In this paper we propose to use the rematch study in a different way than the existing linkage error correction models. Where these existing models first estimate linkage error probabilities and use these probabilities to correct the DS estimate, we directly correct the cell counts in the contingency table for linkage errors. In this way linkage error correction is integrated in the general framework of CR estimation. A cell count represents the size of a group in the combined source, where a group is defined by its source(s). This linkage error corrected contingency table may include multiple sources and covariates and underlies the CR model. Using the log-linear Poisson regression model, statistical inference on this table can be accomplished in the same way as in this model without linkage errors. In this way we derive a CR estimation procedure that corrects for linkage errors but can deal with any number of linked sources and covariates. In Section 4.2 we introduce some notation and discuss the general problem of linkage errors in CR models. In section 4.3 we

first discuss CR models in general and corresponding linkage error correction methods known in the literature. In the same section we combine these to derive a general CR model framework that corrects for linkage errors and can deal with covariates and multiple sources. We refer to this model as the weighted multiple - recapture (WMR) model. The expression *weighted* comes from the individual record weights that we will introduce in Section 4.4. Section 4.5 presents a simulation study that shows that the model works, and section 4.6 concludes and discusses the results.

4.2 Notation and an illustration of linkage errors

In this section we introduce the notation that we use to describe our model. Because our model involves linkage errors, we first discuss source linkage. Imagine there is some linkage procedure ℓ that links a set of sources with linkage keys. A linkage key can either be a perfect identifier γ, like a flawless ID – number, or some set of Z imperfect identifiers $\tilde{\gamma} = \tilde{\gamma}_1, \ldots, \tilde{\gamma}_Z$, such as non – unique names or names that are not spelled flawlessly. In case γ is available, linkage can be performed without linkage errors and in case $\tilde{\gamma}$ is available, source linkage might contain errors. In case more than two sources are available sources can be linked either simultaneously, pairwise, or sequentially. Simultaneously means that all sources are linked in one step. Pairwise means that different sets of sources are linked first after which these linked sources are linked again until all sources are linked into one linked source. Sequential linkage means that first two sources are linked, then the next is linked to this source, and so on, until no sources remain. Each step of sequential linkage could be considered a special version of pairwise linkage. In case γ is available, there is no difference between simultaneous, pairwise, and sequential linkage, they lead to the same result. However, in case only $\tilde{\gamma}$ is available this equality does not necessarily hold. For instance, in case of pairwise linkage, records might be linked inconsistently (e.g. $A \rightarrow B$, $B \rightarrow C$, $C \nrightarrow A$). This inconsistency is not possible in simultaneous or sequential linkage. However, simultaneous linkage has the problem that it can become computationally very intensive, because the number of potential matches increases exponentially with every source. Therefore, as was also argued by DC&T_18, sequential linkage is usually preferred in practice. Therefore, we assume source linkage by ℓ is performed sequentially.

4.2.1 Linkage with perfect identifiers

We first discuss the situation for perfect identifiers. Let there be K sources S^k ($k = 1, \ldots, K$). Each source S^k contains s^k records that represent a set of population units. We assume that the units in each source are a subset of units from the population that has unknown size m. We assume that each source contains a perfect matching key γ that can be used in the (sequential) linkage procedure ℓ. ℓ starts with linking S^1 and S^2 and so on until S^k is linked, which implies a total of $K-1$ linkages. After each step,

the resulting linked source is referred to as N^k after each step. This can be written as:

$$N^k = \begin{cases} N^1 = S^1 \\ N^2 = \ell\left(N^1, S^2 \mid \gamma\right) \\ \cdots \\ N^K = \ell\left(N^{K-1}, S^K \mid \gamma\right) \end{cases} \tag{4.1}$$

where N^k consists of n^k records with $n^k < m$.

4.2.2 Linkage without perfect identifiers

When instead of a perfect, an imperfect linkage key $\tilde{\gamma}$ is available, linkage errors can occur. The number of linkage errors may be reduced with probabilistic linkage models (see e.g. Fellegi & Sunter, 1969; Winkler, 1988; Jaro, 1989). Probabilistic linkage models generally use imperfect linkage keys to minimise both the probability of mismatches and missed matches and find the optimal balance between these two. They estimate, for each possible pair of records, a probability of this pair being a match. For example, when two records have almost the same and unique name, the probabilistic linkage model estimates this pair to have a high probability of being a match and links them. The concepts behind these estimated probabilities will be discussed in more detail in Section 4.3, because they are at the base of the D&F model and its successors. We defined N^k as the combined source that is obtained in case a perfect linkage key γ is available. With the imperfect linkage key $\tilde{\gamma}$, N^k is replaced by $\tilde{N}^k$ and can be written as:

$$\tilde{N}^k = \begin{cases} \tilde{N}^1 = S^1 \\ \tilde{N}^2 = \ell(\tilde{N}^1, S^2 \mid \tilde{\gamma}^1) \\ \cdots \\ \tilde{N}^K = \ell(\tilde{N}^{K-1}, S^K \mid \tilde{\gamma}^K) \end{cases} \tag{4.2}$$

where $\tilde{\gamma}^k$ refers to the set of imperfect linkage key variables that is available in linkage k and $\tilde{N}^k$ has $\tilde{n}^k$ records and may contain mis- and/or missed matches. While with perfect linkage it is certain that the number of records n^k is a lower bound for the population size m, this does not hold for $\tilde{n}^k$. Due to imperfect linkage, $\tilde{n}^k$ can be smaller, equal to or larger than m, but also smaller, equal to or larger than n^k, because a missed match increases the number of records and a mismatch decreases the number of records in $\tilde{N}^k$.

4.2.3 Records and cell counts

N^k and $\tilde{N}^k$ are combined sources with n^k and $\tilde{n}^k$ records, where due to linkage errors a record may represent multiple individuals. A single record r is referred to as N^k_r

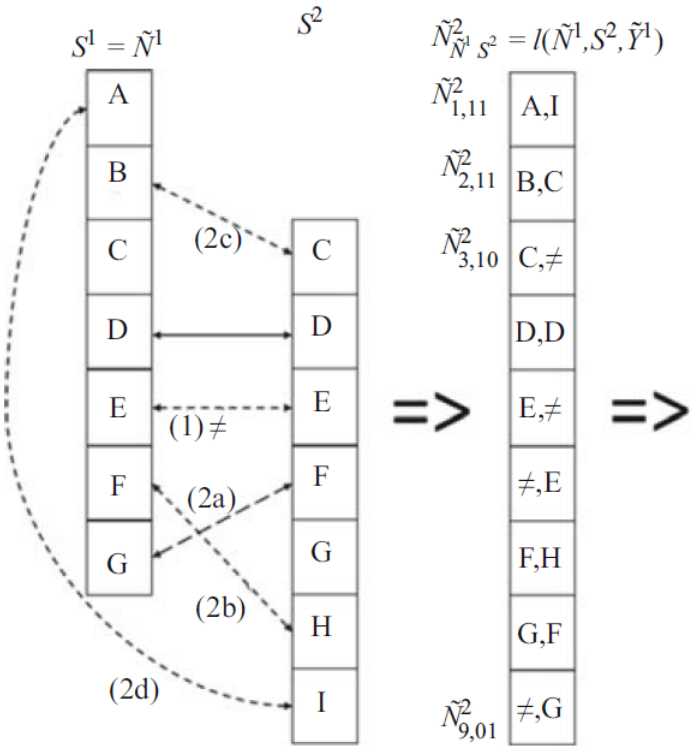
and $\tilde{N}_r^k$ with $r = 1,\ldots,n^k$ and $r = 1,\ldots,\tilde{n}^k$ respectively. Each of these records contains a string in the subscript of K binary indicators that correspond to the sources $S^1 \ldots S^k$ and indicate in which source a record occurs. A frequency then becomes $N^k_{S^1\ldots S^k}$, where e.g. $N^{k=2}_{S^1S^2} = N^{k=2}_{11}$ means the subset of records r that are in both S^1 and S^2. Each subset $N^k_{S^1\ldots S^k}$ has a corresponding cell count denoted as $n_{S^1\ldots S^k}$, which is simply the number of records in subset $N^k_{S^1\ldots S^k}$. These binary indicators are a fundamental part of CR models because they define the cell count categories in the contingency table and serve as explanatory variables. Corresponding with the combined sources N^k and $\tilde{N}^k$ we define A^k and $\tilde{A}^k$ as a matrix with in each row a unique capture history that corresponds to a sum of the observed cell counts $n^k_{S^1\ldots S^k}$ for perfect linkage and $\tilde{n}^k_{S^1\ldots S^k}$ for imperfect linkage. For example, under perfect linkage the unique set of capture histories for three sources is collected in $A^3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ where each row corresponds to an observed count $n_{S^1S^2S^3}$, where the subscripts S^1, S^2 and S^3 refer to a row in A^3. For instance, n_{111} is the count that belongs to the first row in A^3. The observed cell counts $n_{S^1S^2S^3}$ can be considered realisations of a random process, so they also have an expectation that we refer to as $m_{S^1S^2S^3}$. For $m_{S^1S^2S^3}$ we have the equality $\sum_{S^1S^2S^3} m_{S^1S^2S^3} + m_{000} = m$, where m_{000} is the expected number of units in the population missed by S^1, S^2 and S^3. Estimates of $m_{S^1S^2S^3}$, m_{000} and m based on counts resulting from linkage with perfect identifiers $n_{S^1S^2S^3}$, are denoted with a hat, for example $\hat{m}_{S^1S^2S^3}$, while estimates that are based on counts resulting from linkage with imperfect identifiers $\tilde{n}_{S^1S^2S^3}$, are denoted with a reversed hat, for example $\check{m}_{S^1S^2S^3}$.

Finally we note that the definition of A^k above allows for a straightforward extension when categorical covariates are to be included in the process, by adding dummy variables as columns and adding rows such that $S^1 \ldots S^k$ are represented separately for the distinct levels of the covariates. Interactions between the sources, and between sources and covariates, can be included by adding columns appropriately.

4.2.4 An illustration of source linkage, linkage errors and the contingency table

Figure 4.1 illustrates the simple case of the linkage of $k = 2$ sources with the imperfect linkage key $\tilde{\gamma}^1$ and the five linkage errors types (1 – 2d) discussed in section 4.1. The illustration in Figure 4.1 presents two imperfectly linked sources of equal size $s^1 = s^2 = 7$. The total number of units in S^1 or S^2 is nine, and the units are labelled A to I. The solid line arrow represents a correct record linkage while the dashed line arrows represent five other linkages that all correspond to one of the linkage error

Figure 4.1: Illustration of linkage of two sources and different types of linkage errors.



Contingency table			
Linkage cells A^2		Cell counts $n_{S^1S^2}$	$\tilde{n}_{S^1S^2}$
1	1	$n_{11} = 5$	$\tilde{n}_{11} = 5$
1	0	$n_{10} = 2$	$\tilde{n}_{10} = 2$
0	1	$n_{01} = 2$	$\tilde{n}_{01} = 2$

types (1 – 2d). The resulting combined source $\tilde{N}^{k=2}$ contains the nine records $\tilde{N}^{k=2}_r$ ($r = 1,\ldots,9$) and each record belongs to one of the subsets $\tilde{N}^2_{S^1S^2}$. Under perfect linkage each record in $N^{k=2}$ should correspond to one unique unit in S^1 and S^2. This does not hold in case of linkage errors. In fact, in this artificial example the only correct match is [D, D] while all other records represent missed or mismatches. Despite the linkage errors, in this case it (coincidentally) does not lead to errors in the cell counts. The reason is that in this artificial example the five different linkage error types cancel each other out. Obviously, ignoring linkage errors generally lead to a difference between $n_{S^1S^2S^3}$ and $\tilde{n}_{S^1S^2S^3}$. The question we deal with in section 4.4.3 is how we can correct $\tilde{n}_{S^1S^2S^3}$ in such a way that this correction is an unbiased estimate of $n_{S^1S^2S^3}$. But to see why this is useful we first discuss CR models in Section 4.3.

4.3 Linkage error correction in capture - recapture estimation

In this section we describe and discuss DS models and the linkage error correction method introduced by D&F. We first describe the most basic DS model which was introduced by Petersen (1896) and is also known as the Lincoln - Petersen model (Lincoln, 1930). Next, we show how D&F improve this model so that it corrects

for linkage errors. We further discuss DC&T_15, DC&T_18 and WLZ, because they provide the tools that help us to show why correction of the contingency table also corrects for linkage errors.

4.3.1 Relation between the basic dual – system and the log – linear Poisson regression model

In the DS model A^2 has three rows, with associated expected cell counts. The maximum likelihood (ML) estimates $(\hat{m}_{11}, \hat{m}_{10}, \hat{m}_{01})$ are equal to (n_{11}, n_{10}, n_{01}) because the DS model is saturated. Under the appropriate assumptions (Wolter, 1986), including perfect linkage, the basic DS estimate can be obtained by:

$$\hat{m}_{DS} = \hat{m}_{11} + \hat{m}_{10} + \hat{m}_{01} + \hat{m}_{00} = n_{11} + n_{10} + n_{01} + n_{10}n_{01}/n_{11} = s^1 s^2/n_{11} \tag{4.3}$$

where $\hat{m}_{00}$ represents an estimate of the unobserved part of the population and $\hat{m}_{\mathrm{DS}}$ is the estimate for the population size. The expression $s^1 s^2/n_{11}$ simply follows from $s^1 = n_{11}+n_{10}$ and $s^2 = n_{11}+n_{01}$. This expression will become important, because it contains only one value (n_{11}) that can be affected by linkage errors because s^1 and s^2 are simply the size of S^1 and S^2, which are unaffected by linkage errors. The population size can also be estimated using the log-linear Poisson model (see e.g. Cormack, 1989). The log – linear Poisson regression model for A^2 can be written as:

$$m_{S^1S^2} = e^{(\beta_0+\beta_{1,S_1}+\beta_{2,S_2})}, \tag{4.4}$$

with $\beta_{1,S_1}, \beta_{2,S_2} = 0$ if S_1, S_2 in the subscript is zero and a parameter to be estimated otherwise. Using the estimate of the intercept an estimate $\hat{m}_{00}$ can be obtained as $\hat{m}_{00} = e^{\hat{\beta}_0}$. Because the ML estimate $\hat{\beta}_0$ in Eq. (4.3) is $\hat{\beta}_0 = \log(n_{10}) + \log(n_{01}) - \log(n_{11})$, the equality $e^{\hat{\beta}_0} = n_{10}n_{01}/n_{11}$ also holds. This equality shows why Eq. (4.3) and (4.4) lead to the same result. However, an important advantage of the log – linear formulation is that it can be easily extended with additional sources or categorical covariates and the interaction between them. For instance, with a third source and a categorical covariate X with levels 1 and 0, then $m_{S^1S^2}$ becomes $m_{S^1S^2S^3X}$ and the model might for instance be:

$$m_{S^1S^2S^3X} = e^{(\beta_0+\beta_{1,S^1}+\beta_{2,S^1}+\beta_{3,S^1}+\beta_{4,S^1S^2}+\beta_{5,S^1S^3}+\beta_{6,X})}.$$

Extending the Petersen formula in this way would be non - trivial at best, while for each category in X a PSE of the unobserved population can be obtained by $\hat{m}_{0000} = e^{\hat{\beta}_0}$ and $\hat{m}_{0001} = e^{\hat{\beta}_0+\hat{\beta}_{6,X=1}}$.

4.3.2 Impact of linkage errors on the dual - system model

We provide a simple numerical example that illustrates the problem of linkage errors in the DS model. We take $s^1 = 300$, $s^2 = 150$ and $n_{11} = 100$. Due to linkage errors

$\tilde{n}_{11} = 90$. The difference between n_{11} and $\tilde{n}_{11}$ implies that the number of missed links is 10 more than the number of false links. This simple case is represented in Table 4.1 below.

An estimate for m_{00}, $\hat{m}_{00} = n_{10}n_{01}/n_{11} = (200 \times 50)/100 = 100$. However, due to linkage errors not $n_{S_1S_2}$, but $\tilde{n}_{S_1S_2}$ is observed and when this is naively ignored the DS estimate becomes: $\check{m}_{00} = \tilde{n}_{10}\tilde{n}_{01}/\tilde{n}_{11} = (210 \times 60)/90 = 140$, leading to a linkage error bias of 40, something better not left ignored. Note that the "$\vee$" on $\check{m}_{00}$ only means that $\check{m}_{00}$ is an estimate based on cell counts that are subject to linkage errors, not whether the estimate is biased or not.

4.4 The D&F and D&F+ model

The D&F model is a DS model that aims to correct the population size estimate for linkage errors type (1) and (2d) (compare Section 4.1). We refer to the population size estimate resulting from this model as $\hat{m}_{\text{D\&F}}$. To estimate the linkage error probabilities of these two error types, they use a rematch study. A rematch study aims to confirm whether a subset of matches and non – matches were correct or not. The rematch study can be summarized as in Table 4.2.

Table 4.1: Example of true and observed cell counts table of two sources.

A^2		$n_{S^1S^2}$	$n_{S^1S^2}$
1	1	100	90
1	0	200	210
0	1	50	60

Table 4.2: Rematch study with D&F structure.

		Rematch study	
		Matched	Not matched
Probabilistic Linkage	Matched	a_{11}	a_{10}
	Not matched	a_{01}	a_{00}

In Table 4.2 we see how many records in the rematch study were correctly matched (a_{11}), correctly not matched (a_{00}), incorrectly matched (a_{10}) and incorrectly not matched (a_{01}). They define the probability of linkage error type (1) by α and of type (2d) by θ. Thus, $\alpha = a_{11}/(a_{11} + a_{01})$ and $\theta = a_{10}/(a_{10} + a_{00})$ and D&F show how to use these probabilities to obtain $\check{m}_{\text{D\&F}}$ that corrects for linkage errors (1) and (2d). The D&F model recently received more attention from DC&T_15 and WLZ. DC&T_15 write $\check{m}_{\text{D\&F}}$ as:

$$\check{m}_{\text{D\&F}} = (\tilde{n}_{11} + \tilde{n}_{10} + \tilde{n}_{01})/(\hat{p}_1 + \hat{p}_2 - (\alpha - \theta)\hat{p}_1\hat{p}_2 - \theta\hat{p}_1), \tag{4.5}$$

with $\hat{p}_1 = (-N_{11} + \theta(\tilde{n}_{11} + \tilde{n}_{10}))/(\theta - \alpha)(\tilde{n}_{11} + \tilde{n}_{01})$, and $\hat{p}_2 = (-\tilde{n}_{11} + \theta(\tilde{n}_{11} + \tilde{n}_{10}))/(\theta - \alpha)(\tilde{n}_{11} + \tilde{n}_{10})$. These equations show that the D&F model is complex and hard to interpret. The formulas become even more complex when DC&T_15 introduce their so called two - way linkage errors. This model is further extended by WLZ, who show

that in the calculation of the two - way linkage errors it is implicitly assumed that the sizes of source 1 and 2, s^1 and s^2, are equal. Therefore, they extend the model with asymmetrical two – way errors, which allows the value of s^1 and s^2 to differ from each other. Unfortunately, this implies introducing more notational complexity (as θ is separated into θ_1 and θ_2). In DC&T_18 the linkage error correction model is extended from two to three sources. DC&T_18 introduce a so – called transition matrix that allows one to transform the observed cell counts into estimates of the true cell counts, which can serve as input for the Poisson regression model. This is a useful extension on their earlier model, but it is still limited in the sense that the method is not generic with respect to covariates and it is unclear how to add yet an additional source.

Beside WLZ's asymmetrical two – way errors extension, they provide us with another useful contribution. They show that the D&F model, the DC&T_15 model and their own extension all give identical outcomes when not only the formula of $\hat{m}_{\text{D\&F}}$ but also of α and θ are chosen appropriately. They also show that in this case the model corrects for all five linkage error types introduced in Section 4.1. We refer to this model as the D&F+ model. WLZ also show that this can be written much more comprehensively as:

$$\check{m}_{\text{D\&F+}} = s^1 s^2 / \check{n}_{11}, \tag{4.6}$$

where $\check{n}_{11}$ is an estimate for m_{11} based on $\tilde{n}_{11}$ and the rematch study, instead of the directly observed n_{11} used in the DS model. Eq. (4.6) shows that the models derived in D&F, DC&T_15 and WLZ are all equal and a generalisation of the DS-estimator. In the next section we will show that $\check{n}_{11}$ can be derived in a straightforward way when the rematch study is used in a slightly different way. Unlike in WLZ it will no longer depend on α and θ altogether.

4.4.1 Further simplification of the D&F+ model

The D&F+ model as defined in Eq. (4.6) contains only one element that is susceptible to linkage errors, i.e. $\check{n}_{11}$. WLZ derive $\check{n}_{11}$ starting with the a's in table 4.2. These are used to estimate α and θ that in turn are used to estimate $\check{n}_{11}$. In this section we propose to simplify this procedure by using the rematch study differently. To describe this procedure, we first define $\check{N}^k$, which is similar to N^k and $\tilde{N}^k$ as defined in Eq. (4.1) and (4.2) respectively, but with the difference that $\check{N}^k$ is also based on the rematch study in a way that we describe below. The rematch study concerns a representative subsample of the population of which the matches and non-matches were clerically reviewed. This means that for the records in this rematch study subset it is quite simple to count the number of matches before and after clerical review. We refer to the set of records that are subject to clerical review with a "$*$". This implies $\tilde{N}^{k^*}$ and $\check{N}^{k^*}$ are the sets of linked records between $\tilde{N}^{k-1}$ and S^k, that were under clerical review, before and after clerical review. The overlap count of the records in

the clerical review study before and after clerical review are denoted as $\tilde{n}_{11}^{k^*}$ and $\check{n}_{11}^{k^*}$. Then the ratio $\check{n}_{11}^{k^*}/\tilde{n}_{11}^{k^*}$ can be used to estimate $\check{n}_{11}$ with:

$$\check{n}_{11} = \tilde{n}_{11}\check{n}_{11}^{k^*}/\tilde{n}_{11}^{k^*}, \tag{4.7a}$$

For $k = 2$ the elements $\check{n}_{10}$ and $\check{n}_{01}$ can be obtained by:

$$\check{n}_{10} = s^1 - \check{n}_{11}, \tag{4.7b}$$

$$\check{n}_{01} = s^2 - \check{n}_{11}, \tag{4.7c}$$

Note that we write $\check{n}_{11}^{k^*}$ instead of $n_{11}^{k^*}$, although for $k = 2$ they are equal. As we will see later, for $k > 2$ this equality no longer holds, because then $n_{11}^{k^*}$ is no longer a simple count but a sum of weights unequal to 1. Equations (4.7) serve as input for the saturated model as defined in (4.3), i.e. $\check{n}_{S^1S^2} = e^{(\check{\beta}_0+\check{\beta}_{1,S^1}+\check{\beta}_{2,S^2})}$, which gives $\check{m}_{00} = e^{\check{\beta}_0}$. This implies that by combining (4.3) with (4.7) in the basic DS model we have obtained the PSE $\check{m}_{\text{D\&F+}}$, with a simple set of formulas. In the next section we show how these formulas can be extended such that they can deal with covariates and additional sources.

4.4.2 Covariates in the D&F+ model

We proceed by a further development of DS model in the context of the log – linear Poisson regression model with categorical covariates. When there is only one categorical covariate X with $X \in (0, 1)$, then n_{110} is the number of records in S^1 and S^2 with $X = 0$. Note that while without covariates we had $n_{10} = S^1 - n_{11}$, with covariates this can be replaced by $n_{10X} = s_X^1 - n_{11X}$ where s_X^1 refers to the number of records in S^1 for each level in X. This gives us a straightforward way to incorporate covariates in the D&F+ model, because we can simply replace the subscript S^1S^2 in equation (4.7) with the subscript S^1S^2X, which gives:

$$\check{n}_{11X} = \tilde{n}_{11X}\check{n}_{11X}^{k^*}/\tilde{n}_{11X}^{k^*}, \tag{4.8a}$$

$$\check{n}_{10X} = s_X^1 - \check{n}_{11X}, \tag{4.8b}$$

$$\check{n}_{01X} = s_X^2 - \check{n}_{11X}. \tag{4.8c}$$

Equations (4.8) yield $(\check{n}_{11X}, \check{n}_{10X}, \check{n}_{01X})$ that can be used as values of the dependent variable in the log – linear Poisson regression model that includes the covariate X as explanatory variable. This can be extended in a straightforward way for more explanatory variables, as was described in Section 4.3.1. This approach has the advantage that it allows for parsimonious models. For example, it may turn out that some parameters that estimate the effect of covariates do not depart significantly from zero and the model can therefore further ignore this covariate. This option of hypothesis testing is an important improvement over the D&F+ model. Working with a saturated model will induce redundant noise in the DS model, when a more parsimonious

model fits adequately. Therefore, significance testing of covariates is important, and becomes increasingly so when the number of covariates in the CR model increases. Without discussing technical details, we elaborate on the role of X. It is important to include X in the CR model when the capture probabilities are heterogeneous over S^1 and S^2, and X takes this into account. However, it is not necessarily the case that the levels of X differ with respect to linkage error probabilities as well. For instance, records with $X = 1$ might be more likely to be in S^1, however, they are not necessarily also more likely to be falsely linked or not linked to S^2. In this case, despite the significance of X in the CR model, the ratios $n^{k^*}_{S^1S^2X=1}/\tilde{n}^{k^*}_{S^1S^2X=1}$ and $n^{k^*}_{S^1S^2X=1}/\tilde{n}^{k^*}_{S^1S^2X=0}$ should in most cases not differ significantly and X can be ignored in the linkage error correction step. The cell counts of records with $X = 1$ and $X = 0$ can both be corrected with the same ratio $n^{k^*}_{S^1S^2}/\tilde{n}^{k^*}_{S^1S^2}$. Therefore, in practice one may first test whether the ratios $n^{k^*}_{11X}/\tilde{n}^{k^*}_{11X}$ differ significantly from each other for different levels within X.

4.4.3 Additional sources in the D&F+ model: The weighted multiple-recapture model

Eq. (4.7) can be applied on the contingency table of the combined source $\tilde{N}^{k=2}$ (this also holds for (4.8), but we further ignore this to keep the presentation simple). When a third source is involved, it must be linked to $\tilde{N}^{k=2}$ again. However, $\tilde{N}^{k=2}$ was not affected by (4.7), so simply linking S^3 to $\tilde{N}^{k=2}$ would ignore the linkage error correction in (4.7). Therefore, before the next source is linked, the information obtained in this linkage error correction step should somehow be transferred to $\tilde{N}^k$. A straightforward way to do this is by introducing record level weights, which is achieved by disaggregating $\check{n}_{S^1S^2}$ to the record level by distributing $\check{n}_{S^1S^2}$ evenly over the corresponding records. For example for $k = 2$, each record $\tilde{N}^{k=2}_r$ in $\tilde{N}^{k=2}$ receives a weight $w^{k=2}_r = \check{n}_{S^1S^2}/\tilde{n}_{S^1S^2}$ for $r \in \tilde{N}^{k=2}_{S_1S_2}$. We refer to the combination of $\tilde{N}^{k=2}$ and the corresponding vector of linkage error correction weights $w^{k=2}$ as $\check{N}^{k=2}$. $\check{N}^{k=2}$ may now be linked to S^3, giving $\tilde{N}^{k=3}$, which may introduce new linkage errors. $\tilde{N}^{k=3}$ can be used to obtain $\hat{m}_{\check{N}^{k=2}S^3}$ by summing up over $w^{k=2}_r$ for the records in $\check{N}^{k=2}$ while (new) records in S^3 receive a weight $w^{k=2}_r = 1$. This gives cell counts that are corrected for linkage errors in going from S^1 to S^2 but not yet in going from $\check{N}^{k=2}$ to S^3. To correct for these new linkage errors the linkage error correction step in (4.7) can be repeated to transform $\hat{m}_{\check{N}^{k=2}S^3}$ into $\check{m}_{\check{N}^{k=2}S^3}$. In case more sources are linked, this linkage error correction procedure can be repeated after each new source. This procedure of linking two sources, aggregating this combined source to a contingency table, correcting the cell counts for linkage errors, disaggregation the contingency table back to the combined source and again linking a new source, is quite cumbersome. This procedure becomes more straightforward when the linkage error correction step in (4.7) is performed directly on the record level weights w^k_r. Then, only after the last source is linked, a contingency table that is corrected for linkage errors is produced by summing up over the weights for the corresponding categories. This can be written more

formally by an updating scheme for w_r^k with $w_r^{k=1} = 1$:

$$w_r^k = \begin{cases} w_r^{k-1}\check{n}_{11}^{k^*}/\tilde{n}_{11}^{k^*} & \text{for } r \in \tilde{N}_{11}^k \\ w_r^{k-1}\check{n}_{10}^{k^*}/\tilde{n}_{10}^{k^*} = \left(\check{n}^{(k-1)^*} - \check{n}_{11}^{k^*}\right)/\left(\check{n}^{(k-1)^*} - \tilde{n}_{11}^{k^*}\right) & \text{for } r \in \tilde{N}_{10}^k \\ 1 \qquad \check{n}_{01}^{k^*}/\tilde{n}_{01}^{k^*} = \left(s^{k^*} - \check{n}_{11}^{k^*}\right)/\left(s^{k^*} - \tilde{n}_{11}^{k^*}\right) & \text{for } r \in \tilde{N}_{01}^k. \end{cases} \tag{4.9}$$

Where $s^{k^*} = \sum_{r\in(\check{N}_{11}^{k^*},\check{N}_{01}^{k^*})} w_r^{k-1}$, $\check{n}^{(k-1)^*} = \sum_{r\in(\check{N}_{11}^{k^*},\check{N}_{10}^{k^*})} w_r^{k-1}$, $\check{n}_{11}^{k^*} = \sum_{r\in\check{N}_{11}^{k^*}} w_r^{k-1}$ and $\tilde{n}_{11}^{k^*} = \sum_{(r\in\tilde{N}_{11}^{k^*})} w_r^{k-1}$. Note that records with $r \in \tilde{N}_{01}^k$ are always new records that were not linked in the $k-1$ previous linkage steps. Therefore, their (individual starting) weight is simply (still) equal to 1, because they were not updated in any of the previous updating steps. Further note that in case there is reason to believe some covariate groups may be more susceptible to linkage errors than others, Eq. (4.9) may be applied for these groups separately. Generally, the record level linkage error correction weight w_r^k is a weight that can be interpreted in a similar way as well-known individual sample weights in survey models. In survey models, individual sample weights allow a researcher to correct for over- and underrepresentation of specific groups in a survey. A record with a higher than average weight belongs to a group that is underrepresented and vice versa for a record with a low weight. Similarly, a record with a higher or lower than average linkage error correction weight belongs to a group with a cell count that is under- or overestimated, respectively. With individual sample weights, it is quite common to sum up over these weights to obtain representative totals. For instance, when the number of men is underrepresented, summing up over their sample weights gives the number of men that is corrected for this underrepresentation. The same reasoning holds for the record level linkage error correction weights. By applying Eq. (4.9) after each source linkage, a contingency table that is corrected for linkage errors can be constructed after every source linkage. This contingency table is different from a regular contingency table that simply counts the number of records in each linkage cell. The linkage error corrected contingency table is constructed by summing up the weights of these records over these linkage cells instead of counting records. Therefore, we refer to the models based on this contingency table as the weighted dual – system (WDS) model for two sources and the weighted multiple – recapture (WMR) model for more than two sources. In case there are no linkage errors, the models reduce to the standard DS and MR models.

4.5 Simulation study

We evaluate the WMR model with a simulation study. The main goal of this simulation study is to study whether our new WDS and WMR model behave under different conditions, such as (no) linkage errors, (no) covariate dependence, (no) source dependence and combinations thereof. In Section 4.5.1 we describe the setup of this

simulation study and in section 4.5.2 we discuss the results.

4.5.1 Simulation study setup

For the simulation study to reflect reality as well as possible we use a quasi - real dataset that is publicly available and represents a fictitious population dataset of 26,625 persons. It is constructed such that it is representative for the UK census population. It was created in the ESSnet DI (McLeod et al., 2011), a European project on data integration (Record Linkage, Statistical Linking, Micro integration Processing) that ran from 2009 to 2011. The dataset has linkage keys such as address and birthdate but also covariates such as gender and age. In each replication of the simulation study a random population of 10,000 is generated. This size of 10,000 is chosen because the Poisson regression estimators have known finite sample bias (see e.g. Chapman, 1951; Menkens & Anderson, 1988; Q. Chen & Giles, 2011). This bias goes to zero when the sample increases to infinity. For, say, a population size of 1,000 this bias may still play a role, so then it will be hard to say whether a CR model corrects for linkage error bias. A probable example of this finite sample bias can be found in DC&T_18 who present a simulation study with similar data and setup but with a true population size (TPS) of 1,000. In this study, the mean of the PSEs that were unaffected by linkage errors deviates slightly but statistically significantly (i.e. by 1.05%) from the TPS. This small bias is like the finite sample bias that we encountered when we experimented with a TPS of 1,000. Unfortunately, the population size can also not be too large because probabilistic record linkage is computationally very intensive. A population size of 10,000 is a balanced choice that leaves the finite sample bias practically ignorable and leaves the probabilistic linkage procedure computationally feasible. This population of 10,000 serves to generate three sources that each cover part of the population. These sources are generated under different conditions where conditions vary with respect to covariate and source dependence. This leads to four scenarios:

1. Three randomly generated sources (no dependencies).
2. Three sources in which covariates affect the probability of a record to be in a source (covariate dependence).
3. Three sources where the probability of a record to be in a source is affected by this record being in other sources (source dependence).
4. Three sources where records are subject to both covariate and source dependence.

Next, in each replication the sources are linked both with and without linkage errors. The linked sources allow us to apply both the regular (referred to as naïve) and weighted DS (using only the first two sources) and MR (using all three sources) model. By replicating this procedure many times (i.e. 1,050[1] for each scenario) we can obtain a distribution of estimates that in case the model provides asymptotically unbiased estimates will evolve around the TPS of 10,000. In this way we can see whether the WDS and WMR model can deal with covariate and source dependencies while suffering from linkage errors, conditions under which the regular DS and MR model fail. A more detailed description of the simulation setup can be found in Appendix 4.7.2.

4.5.2 Simulation results

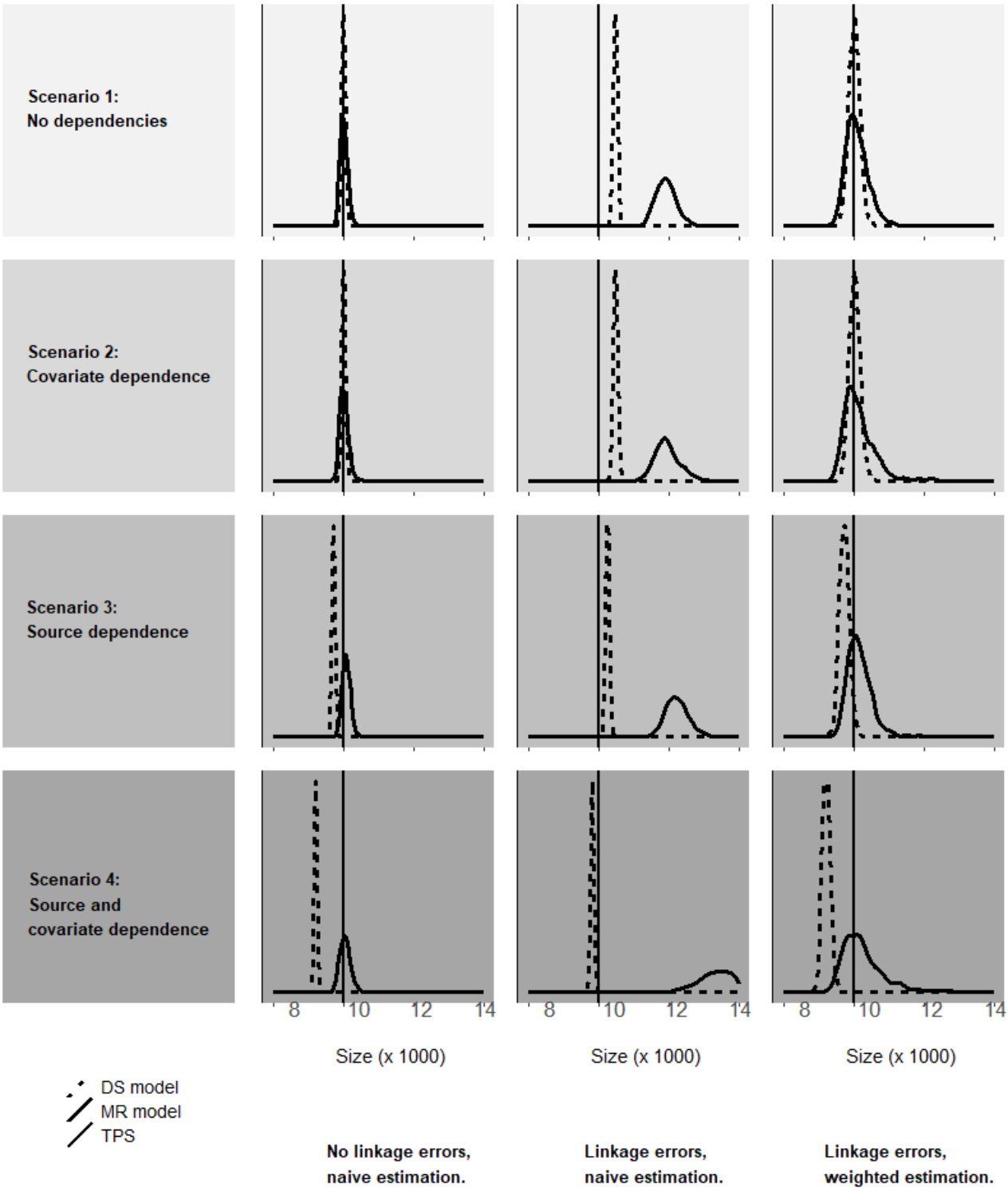
In Figure 4.2 below the simulation results of the four scenarios are presented as density plots. Figure 4.2 contains twelve density plots that each contain distributions/-densities of DS and MR estimates. In the rows there are the four scenarios, and the first two columns distinguish naïve estimation with and without linkage errors. The third column shows the WDS and WMR model in case of linkage errors. The graph clearly shows that the estimates that can be expected to be biased, are biased. However, most importantly, it shows that in case of linkage errors the weighted estimates are on target while naïve estimates are biased. Furthermore, the presence of covariate dependence is no problem for the weighted estimates, even in combination with source dependence. A numerical calculation example of one of the replications can be found in Appendix 4.7.1.

4.6 Discussion

In this paper we derived and tested the WMR model for population size estimation corrected for linkage error. The model is derived from the D&F model and is a more general extension than the models developed by Di Consiglio and Tuoto (2015, 2018) and de Wolf et al. (2019) because it includes three or more sources and covariates, which are often necessary to correct for other sources of bias. The linkage error correction model we developed is incorporated in the more general family of log - linear regression models. Thus, linkage error no longer has to be studied as an isolated issue in CR models. Finally, the WMR model was tested and approved in a simulation study. In practise the WMR model does not solve all the linkage error problems. For instance, it still requires a rematch study in which for a share of records clerical review is required to check whether they were correctly linked or not linked. Ideally these records are representative for the records in both sources, both with respect to

[1]The number is 'only' 1,050 because we use a Spark cluster of fifteen cores (available at Statistics Netherlands mainly for Big Data related computations) that each do 70 replications with different random seeds, in which each single replication takes about 10 minutes. In total it took almost two days to run all four scenarios, which is mainly due to the computation time of the probabilistic linking of the three sources.

Figure 4.2: Density plots of two PSEs with three dependent variables and four scenarios.



covariates but also the quality of linkage keys. This last element should not be underestimated. When for instance the records in the rematch study are based on their high – quality linkage keys (which makes clerical review easier), they might suffer less from linkage errors than other records. This will lead to a biased correction. Another issue is the size of the rematch study, when the sources contain some small groups of records, it might be hard to find enough records of this group to perform clerical review. How large the impact of such issues is, requires further research. Also, we should note that we paid little attention to the impact of the exact linkage procedure. We developed the WMR model in the context of sequential linkage, in which first two sources are linked, and a third source is linked to this combined source. We think that in theory the order of linkage does not matter and also pairwise linkage (link each pair and then combine them into one) or simultaneous linkage (link all sources at once) can be incorporated into the WMR model, although this would require further research. In practise the exact linkage strategy may play a role, mainly because linkage is also often used to enrich sources. When, for instance one source contains data on say gender and another on income, the combined source usually contains both, which will probably affect the quality of linkage with a third source that also contains gender and income. Another point that deserves some discussion is the 'individual starting weight of 1'. Lists or registers of individuals sometimes also contain individual sample weights, which indicate the size of the group that this individual represents as part of the total population. The proportion of the sample weights of these new records in relation to the weights of records that were already known from previous records can be used to improve these starting weights. Furthermore, when additional sources also contain sample weights they can be used to construct the cell counts in the contingency table by adding up over weights instead of counting the records. In this way we would get 'linkage error corrected sample weights'. How and when sample weights can be combined with linkage and linkage error correction requires further research.

4.7 Appendix

4.7.1 numerical calculation example

As an illustration of the method we present one of the replications generated under scenario 4 in the simulation study. In Table 4.3 we show the total cell counts with linkage errors together with the audit study cell counts. In the last column we show the correction of groups of individual weights. Table 4.3: Contingency table and correction of weights after linking S^1 and S^2.

Table 4.4 is like Table 4.3 but shows the linkage of $\tilde{N}^2_{S^1S^2}$ and S^3, together with the audit study. The last column shows the second update of weights. Note the relation between the columns $\tilde{n}_{S^1S^2X}$ and w^2_r in Table 4.3, and column $\tilde{n}_{\tilde{N}^2S^3X}$ in Table 4.4, which can be seen by $2784 \times (222/264) + 1080 \times (180/138) + 164 \times (64/12) = 502.2 + 4122.25$ for $X = 0$ and $2030\times(152/226)+2292\times(314/240)+82\times(44/10) = 1789+2235.81$

Table 4.3: Contingency table and correction of weights after linking S^1 and S^2.

Linkage cells		Covariate	Cell counts			Weight correction
S^1	S^2	X	$\tilde{n}_{S^1S^2X}$	$\tilde{n}^{k^*}_{S^1S^2X}$	$n^{k^*}_{S^1S^2X}$	w_r^2
1	1	0	2784	264	222	$r \in \tilde{N}^2_{110} : w_r^2 = 222/264$
1	0	0	1080	138	180	$r \in \tilde{N}^2_{100} : w_r^2 = 180/138$
0	1	0	164	12	64	$r \in \tilde{N}^2_{010} : w_r^2 = 64/12$
1	1	1	2030	226	152	$r \in \tilde{N}^2_{111} : w_r^2 = 152/226$
1	0	1	2292	240	314	$r \in \tilde{N}^2_{101} : w_r^2 = 314/240$
0	1	1	82	10	44	$r \in \tilde{N}^2_{011} : w_r^2 = 44/10$

for $X = 1$

Table 4.4: Contingency table and correction of weights after linking $\tilde{N}^2$ and S^3.

Linkage cells		Covariate	Cell counts			Weight correction
$\tilde{N}^2$	S^3	X	$\tilde{n}_{\tilde{N}^2S^3X}$	$\tilde{n}^{k^*}_{\tilde{N}^2S^3X}$	$n^{k^*}_{\tilde{N}^2S^3X}$	w_r^3
1	1	0	502.2	62.74	36	$r \in \tilde{N}^3_{110} : w_r^3 = w_r^2(36/62.74)$
1	0	0	4122.25	410.34	412	$r \in \tilde{N}^3_{100} : w_r^3 = w_r^2(412/410.34)$
0	1	0	344	38	2	$r \in \tilde{N}^3_{010} : w_r^3 = 1(2/38)$
1	1	1	1789	194.14	174	$r \in \tilde{N}^3_{111} : w_r^3 = w_r^2(174/194.14)$
1	0	1	2935.81	272.39	296	$r \in \tilde{N}^3_{101} : w_r^3 = w_r^2(296/272.39)$
0	1	1	1798	188	14	$r \in \tilde{N}^3_{011} : w_r^3 = 1(14/188)$

Finally, Table 4.5 shows the contingency tables that underly that MR models. In the three rows at the bottom there are the different Poisson regression estimates of the unobserved parts of the population, for $X = 0$ and $X = 1$, together with the total population size estimate.

4.7.2 Setup of the simulation study

From the available dataset we use the file "`person_list.csv`". This list contains both a perfect identifier (id - code) and linkage keys (e.g. surname, address) and can therefore be used to link records both perfectly (i.e. deterministically without any errors) and probabilistically. In this simulation study we use a set of three linkage keys .

Table 4.5: Contingency table and correction of weights after linking S^1, S^2 and S^3.

Linkage cells A^3			Covariate X	Cell counts and sum of weights w_r^3		
				$n_{S^1S^2S^3X}$	$\tilde{n}_{S^1S^2S^3X}$	$\check{n}_{S^1S^2S^3X}$
1	1	1	0	200	264	165.99
1	1	0	0	2328	2440	2060.1
1	0	1	0	82	106	79.34
1	0	0	0	1254	974	1275.56
0	1	1	0	44	14	42.85
0	1	0	0	704	150	803.23
0	0	1	0	18	344	18.11
1	1	1	1	452	766	461.74
1	1	0	1	872	1264	910.45
1	0	1	1	1102	866	1015.47
1	0	0	1	1896	1426	1998.07
0	1	1	1	1896	32	126.19
0	1	0	1	310	50	235.61
0	0	1	1	94	1798	133.89
Total				**9506**	**10574**	**9327**
0	0	0	0	378.34	4774.67	439.70
0	0	0	1	173.25	2803.45	249.46

In order to have a certain degree of linkage errors, in each linkage key in each source, we replace 3% of the records with a random value that can also be found in the population for that linkage key (e.g. replace a surname with a random surname), where in each source, each record has the same probability to be selected. Furthermore, the list contains several covariates, of which we use 'SEX' as covariate X to affect capture probabilities. For each replication first a random population of 10,000 records is generated (without replacement) from the person list. Our aim is then to generate three sources of different sizes from this population (approximately 8,000, 5,000 and 2,000 records) that may suffer from source and covariate dependence. The introduction of source dependence is not straightforward, because source dependence implies that no single source may be independent of other sources. However, when the first source would be generated while other sources do not yet exist, this first source is independent of these other sources. Therefore, before the first source is generated, we first generate three so called latent sources U^k with $k = 1, 2, 3$ of 8,000 units each, which are simply random samples from the population of 10,000. These three latent sources allow us to introduce dependencies between sources such that no source S^k is independent of the other sources. This is done by giving each unit $u = 1, \ldots, 10,000$ a probability to be in each source K by:

$$P_u^k\left[S^k = 1\right] = 1/\left(1 - \exp\left(-\mu_u^k\right)\right), \tag{4.10}$$

where for each population unit u we can write $\mu_u^k = \delta_{U^1}^k U_u^1 + \delta_{U^2}^k U_u^2 + \delta_{U^3}^k U_u^3 + \delta_X^k X_u$. Given Eq. (4.10) we can vary δ's and hereby control dependencies between any source in S^k and the other two sources in S^k and the covariate. For instance, when $\delta_{U^1}^1, \delta_{U^1}^2, \delta_{U^2}^1, \delta_{U^2}^2 \neq 0$, the probability of a record to be in S^1 depends on it being in S^2 while the probability to be in S^2 also depends on it being in S^1. Furthermore, the δ's control the size of each source. The values for the δ's in the simulation study are in Table 4.6.

Because the varying of δ's affects the capture probabilities of units, different δ's also correspond to different estimates of the β's from the Poisson regressions. To assure that by varying δ's we introduce a substantial source and covariate dependence, Table 4.7 presents the (corresponding) mean values of estimated β's over all replications of the benchmark case of no linkage errors.

Table 4.7 clearly shows that the estimated β's correspond to the four scenarios. Each scenario has a column and if for that scenario a statistical significant relation does not exist for a certain parameter, this is indicated by a "×". Statistical significant relations are indicated by a value which is the mean value of the corresponding estimated β for that relation. In scenario 1 neither covariate X nor another source plays a significant role in describing the observed frequencies. In scenario 2 the observed frequencies do not depend on other sources but do depend on X. In scenario 3 the covariate X is not significant while the other sources have significant explanatory power. In scenario 4 both X and the other sources play a significant role. Finally, the last necessary elements of the simulation study are $\check{N}^{2^*}$ and $\check{N}^{3^*}$, which are generated

Table 4.6: Parameter values of the four different scenarios.

Scenario 1	$\delta^k_{U^1}$	$\delta^k_{U^2}$	$\delta^k_{U^3}$	δ_X
μ^1_u	6.3	0	0	0
μ^2_u	0	3.5	0	0
μ^3_u	0	0	1.9	0
Scenario 2	$\delta^k_{U^1}$	$\delta^k_{U^2}$	$\delta^k_{U^3}$	δ_X
μ^1_u	5.6	0	0	2
μ^2_u	0	4.6	0	-2
μ^3_u	0	0	0.42	2
Scenario 3	$\delta^k_{U^1}$	$\delta^k_{U^2}$	$\delta^k_{U^3}$	δ_X
μ^1_u	4.8	1.8	0	0
μ^2_u	0	3.5	0	0
μ^3_u	0	-0.5	2.3	0
Scenario 4	$\delta^k_{U^1}$	$\delta^k_{U^2}$	$\delta^k_{U^3}$	δ_X
μ^1_u	3.9	1.5	0	2
μ^2_u	1.5	3.3	0	-2
μ^3_u	0	-0.5	1.8	1

Table 4.7: Parameter values of the four different scenarios.

Scenario	1^*	2^*	3^*	4^*
Variable$\setminus$Estimate	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
Constant	13.0	12.8	13.0	13.3
S^1	1.3	1.1	1.2	0.3
S^2	×	0.7	-0.2	×
S^3	-1.3	-2.7	-1.3	-1.6
X	×	-0.1	×	-0.6
S^1X	×	0.6	×	1.5
S^2X	×	-1.5	×	-1.9
S^3X	×	2	×	0.8
S^1S^2	×	×	0.4	1.1
S^1S^3	×	×	×	×
S^2S^3	×	×	-0.2	-0.2
S^1S^2X	×	×	×	0.2
S^1S^3X	×	×	×	×
S^2S^3X	×	×	×	0.1

$*$ indicates "scenario without linkage errors"

by first selecting a random 10% (without replacement) of the population and within this selection only keeping those records that are also in S^1 and S^2 (for $\check{N}^{2^*}$) or S^2 and S^3 (for $\check{N}^{3^*}$). We compare three types of PSEs, naïve, perfect, and weighted. Naïve PSEs are estimates based on $\tilde{n}$, so linkage errors are present but ignored. Perfect PSEs are based on n, so linkage errors are not present (and ignored). Weighted PSEs are based on $\check{n}$, so linkage errors are present but if the model works it should correct for them. Finally, for each scenario and PSE type, the DS and MR model are applied.

Chapter 5

From Quarterly to Monthly Turnover Figures Using Nowcasting Methods

Short-term business statistics at Statistics Netherlands are largely based on Value Added Tax (VAT) administrations. Companies may decide to file their tax return on a monthly, quarterly, or annual basis. Most companies file their tax return quarterly. So far, these VAT based short-term business statistics are published with a quarterly frequency as well. In this article we compare different methods to compile monthly figures, even though a major part of these data are observed quarterly. The methods considered to produce a monthly indicator must address two issues. The first issue is to combine a high- and low-frequency series into a single high-frequency series, while both series measure the same phenomenon of the target population. The appropriate method that is designed for this purpose is usually referred to as "benchmarking". The second issue is a missing data problem, because the first and second month of a quarter are published before the corresponding quarterly data are available. A "nowcast" method can be used to estimate these months. The literature on mixed frequency models provides solutions for both problems, sometimes by dealing with them simultaneously. In this article we combine different benchmarking and nowcasting models and evaluate combinations. Our evaluation distinguishes between relatively stable periods and periods during and after a crisis because different approaches might be optimal under these two conditions. We find that during stable periods the so-called Bridge models perform slightly better than the alternatives considered. Until about fifteen months after a crisis, the models that prone heavily on historic patterns such as the Bridge, MIDAS and structural time series models are outperformed by more straightforward (S)ARIMA approaches.

This chapter is published in the Journal of Official Statistics: Zult, D.B. (DZ), Krieg, S. (SK), Schouten, B. (BS), Ouwehand, P. (PO) and van de Brakel, J. (JvdB), 2023. From Quarterly to Monthly Turnover Figures Using Nowcasting Methods. Journal of Official Statistics, Vol. 39, No. 2, 2023, pp. 253–273 https://doi.org/10.2478/jos-2023-0012. Author contributions: The department of business statistics at Statistics Netherlands posed the problem, DZ, SK, PO and JvdB discussed the problem and ideas, BS provided the data for an early version of the article, DZ, PO and BS discussed the non-STM models, SK and JvdB discussed the STM models and SK did the STM calculations. DZ did the other calculations, analyses and wrote most of the text while in regular discussions with SK, who also wrote the text that concerns STM theory. SK, PO and JvdB discussed and edited the text.

5.1 Introduction

The purpose of national statistical institutes (NSIs) is to publish relevant, accurate and timely official statistics. However, the production of high-frequency timely statistics generally compromises the accuracy of these figures. This trade-off is even increased if a major part of the data are observed on a frequency that is lower than the required output. For instance, in The Netherlands, short-term business statistics rely for the most part on turnover obtained from value added tax (VAT) administrations, where most companies declare turnover either quarterly or monthly. The current approach is to wait for the quarter to be finished and produce a quarterly statistic for a PPC, based on turnover data from both monthly and quarterly declarants that are available at the publication date. The question is whether the same data can also be used to produce an earlier, more frequent, and sufficiently accurate monthly estimate at this detailed level. This question has two distinctive elements.

The first element is an increase in frequency, which can be achieved with a method referred to as benchmarking (BM). BM models are extensively discussed in the literature, such as in the "ESS Guidelines on temporal disaggregation, benchmarking, and reconciliation" (Eurostat, 2018). In the case of sufficiently long time series, which is what we assume in this paper, these guidelines recommend BM models that are based on movement preservation as in Denton (1971); Chow and Lin (1971); Dagum and Cholette (1975). We will briefly discuss these BM models in Section 5.3.

The second element is an increase in timeliness, which can be achieved with a method referred to as nowcasting (NC). The most straightforward NC approach would be to extrapolate the monthly series obtained with BM by change in turnover of monthly declarants over the previous and current month. This would closely follow the current production process of the quarterly series and is therefore attractive. However, because the monthly declarants may constitute a selective and/or small sample of the population, this nowcast probably needs to be adjusted as soon as the quarterly data is available. To minimize this adjustment, it is worthwhile investigating some more sophisticated NC models.

NC models are extensively discussed in literature, such as in in the "Handbook of rapid estimates" (Eurostat, 2017). This includes mixed frequency models that combine the BM and NC problem in one model. We discuss NC models in more detail in Section 5.2.3. Simultaneous (multivariate) estimation of series for different PPCs might improve the accuracy of the estimates. We leave this for further research.

Another sample selection and sample size bias correction approach would be to weight the monthly declarants to the entire population with the help of background characteristics and trends. This method is called pseudo design-based estimation (Baker et al., 2010). In our case we deem this approach less fruitful, because the background characteristics that are available are limited (i.e. number of employees). However, maybe by using historic turnovers on the company level in an imputation model, it might be possible to improve our auxiliary monthly series and hereby our nowcasting results. This is outside the scope of this paper but might be worthwhile

further research.

It is important to note that the combination of a BM and NC model introduces an evaluation problem. Normally NC models can be evaluated by simply waiting for the future to unfold, after which a prediction can be compared with a true value. However, because some of the respondents respond on a quarterly basis, the unfolding BM model monthly series may be more accurate than the NC model results, but they are not true values. Which (type of) NC model is best suited for this problem is not trivial and is an important topic of this paper. Furthermore, this question requires elaboration on evaluation criteria, which are therefore discussed in Section 5.2.4.

A final element that deserves special attention, especially in the light of the recent COVID-19 pandemic and resulting lockdowns, is the impact of extreme (economic) developments. Therefore, in Section 5.2.4, we apply and evaluate the BM and NC models discussed in Section 5.2 both in stable and extreme conditions. Both conditions are present in several economic sectors in the Netherlands over the period January 2010 to June 2020, which can be characterized as a long, economically stable period that ends with the COVID-19 pandemic. To further investigate the long term impact of a crisis on the different models, we also simulate different types of crises and evaluate how the accuracy of the nowcasts is affected during and after a crisis. In Section 5.4 we conclude. The supplemental file that belongs to this paper contains some additional technical details about the models discussed in this article.

5.2 Notation benchmarking models and nowcasting models

This section introduces some notation and assumptions (Section 5.2.1) and discusses BM (Section 5.2.2) and NC models (Section 5.2.3). In the main text we discuss the models primarily at the conceptual level with a modest level of technical detail because they are already well described in the literature. We first introduce some notation that allows us to describe the problem and the BM and NC models.

5.2.1 Notation

Let y_t^Q be an observed quarterly time series with $t = 1, 2, \ldots, T_q$ where $t = 1$ is the first month of the index series (e.g. January 2010) and T_q is the third month of the last available quarter. Furthermore, $Q(t)$ is a function that transforms t to the first month of its corresponding quarter. This implies, for example, that when $t = 11$, then $Q(t) = 10$, or when $t = 25$, then $Q(t) = 25$. It is understood that y_t^Q is written as a monthly series, where the months within a quarter are equal, i.e. $y_{Q(t)}^Q = y_{Q(t)+1}^Q = y_{Q(t)+2}^Q$. The three equal elements $y_{Q(t)}^Q$, $y_{Q(t)+1}^Q$ and $y_{Q(t)+2}^Q$ are all observed at time $Q(t)+3$ (i.e. the first month of the next quarter). In the intermediate months, y_t^Q does

not change. Furthermore, there is a monthly auxiliary time series x_t with $t = 1, 2, \ldots, T_m$ where x_t is observed in month t and T_m is the last available month (i.e. $T_q = Q(T_m) - 1$). In other words, the monthly series x_t always extends 1, 2 or 3 months beyond the quarterly series y_t^Q. Next, we define an unobserved monthly time series y_t^M with $t = 1, 2, \ldots, T_{\max}, (T_{\max} \leq T_m)$ which is simply the quarterly series y_t^Q, disaggregated to a monthly series that goes into a yet unobserved future (until $T_{\max}$). The series y_t^Q and y_t^M are related within each quarter by either their mean (e.g. in case of an index) or their sum (e.g. in case of turnover). Therefore, we can write either $y_{Q(t)}^Q = \left(y_{Q(t)}^M + y_{Q(t)+1}^M + y_{Q(t)+2}^M\right)/3$ or $y_{Q(t)}^Q = y_{Q(t)}^M + y_{Q(t)+1}^M + y_{Q(t)+2}^M$ with $t = 1, 2, \ldots, T_q$. Ideally, we would observe y_t^M, but instead we only observe the quarterly aggregate y_t^Q and the related variable x_t. Therefore, we also define $\hat{y}_t^M$ with $t = 1, \ldots, T_m$, which is an estimate of y_t^M based on y_t^Q and x_t. Note that $\hat{y}_t^M$ is the target series of this paper. Because $\hat{y}_t^M$ depends on the available information at the time of estimation, we also define $\hat{y}_{t|T}^M$, which is $\hat{y}_t^M$ estimated given the information available at time T. The appropriate method to estimate $\hat{y}_{t|T}^M$ depends on the data available about time t at time T. When both y_t^Q and x_t are available for t (i.e. for $t \leq T_q$), $\hat{y}_{t|T}^M$ can be estimated with a BM model. We denote this BM estimate by $\hat{y}_{t|T}^{M,\text{BM}}$ where BM indicates the type of BM model. When only x_t is available for time t (i.e. for $T_q < t \leq T_m$), $\hat{y}_{t|T}^M$ should be estimated with a NC model, denoted as $\hat{y}_{t|T}^{M,\text{BM,NC}}$ where NC indicates the NC model. Note that $\hat{y}_{t|T}^{M,\text{BM,NC}}$ also contains BM in the superscript, because a nowcast changes when the target series is the result of a different BM model. The most interesting element in $\hat{y}_t^M$ from a methodological perspective is $\hat{y}_{T_m|T_m}^M$. At this month T_m, x_t with $t = 1, \ldots, T_m$ is known, but $y_{T_m}^Q$ is not. The full series $\hat{y}_{t|T_m}^M$ can be written as:

$$\hat{y}_{t|T_m}^M = \begin{cases} \hat{y}_{t|T_m}^{M,\text{BM}} & \text{for } t = 1, \ldots T_q \\ \hat{y}_{t|T_m}^{M,\text{BM,NC}} & \text{for } t = Q(T_m), \ldots, T_m \end{cases} \tag{5.1}$$

The series $\hat{y}_{t|T_m}^{M,\text{BM}}$ is based on a BM model that uses y_t^Q and x_t with $t = 1, \ldots, T_q$ as input. This implies that each element in $\hat{y}_{t|T_m}^{M,\text{BM}}$ changes each time a new quarter becomes available. The series $\hat{y}_{t|T_m}^{M,\text{BM,NC}}$ is based on both a BM and NC model and uses y_t^Q with $t = 1, \ldots, T_q$ and x_t with $t = 1, \ldots, T_m$ as input, so $\hat{y}_{t|T_m}^{M,\text{BM,NC}}$ changes each time a new month in x_t becomes available.

Finally, some nowcasting models provide a quarterly estimate $\hat{y}_{t|T_m}^{Q,\text{NC}}$ for $t = Q(T_m), \ldots, Q(T_m) + 2$. Therefore we define:

$$\hat{y}_{t|T_m}^M = \begin{cases} \hat{y}_{t|T_m}^{M,\text{BM}} & \text{for } t = 1, \ldots T_q \\ \hat{y}_{t|T_m}^{M,\text{BM,NC}} & \text{for } t = Q(T_m), \ldots, T_m, \end{cases}$$

$$\hat{x}_t = \begin{cases} x_t & \text{for } t = 1, \dots T_q \\ \hat{x}^{\text{NC}}_{t|T_m} & \text{for } t = T_m + 1, \dots, Q(T_m) + 2, \end{cases}$$

which are simply the series y_t^Q and x_t extended with nowcasts for missing values in the current quarter. In this paper, x_t is based on the monthly VAT declarants and y_t^Q is based on the combination of all declarants (monthly and quarterly). Furthermore, both series are assumed to be index series, as this is the publication format of these short-term business statistics.

5.2.2 Benchmarking models

BM models have the aim of temporally disaggregating a low frequency series into a high-frequency series, with the help of (an) auxiliary high-frequency series. BM can be considered a specific case of temporal disaggregation (Eurostat, 2018), where the high-frequency indicator series and the low frequency benchmark series describe the same phenomenon, as is the case in our problem. Extensive literature is available on BM models, see Eurostat (2018) for an overview. The most basic BM model is developed by Denton (1971); Dagum and Cholette (1975, DC) and a slightly more advanced BM model is developed by Chow and Lin (1971, CL). The DC and CL models are widely used in the production of official statistics and are implemented in standard software (Barcellan & Buono, 2002). Both models are also discussed in the "Handbook of Rapid Estimates" (Eurostat, 2017), because they also might be considered as mixed frequency nowcasting models, which will be discussed in the next section. Both models can disaggregate a quarterly series y_t^Q into a monthly series with the help of a monthly auxiliary series x_t for $t = 1, \dots, T_q$, such that both series are consistent in each quarter. Furthermore, both models aim at movement preservation of the high-frequency series x_t. There are also other BM models available, such as by Fernández (1981) and Litterman (1983), but they are better suited for non-stationary residual models, which in our case is less likely because we use two series that measure the same phenomenon. Both the DC and CL model require high and low-frequency data over the same period. This implies that a new monthly benchmarked estimate $\hat{y}^{M,\text{BM}}_{t|T_m}$ can be obtained only each time new quarterly data becomes available.

The difference between CL and DC is that DC aims to preserve the movement by mimicking the month-on-month growth in x_t as close as possible (minimising either the proportional or absolute deviations), while CL is a regression approach that controls for the estimated relation between y_t^Q and x_t. Furthermore, the CL model can deal with more than one auxiliary time series. When the pattern in x_t is representative for the pattern in y_t^Q, DC and CL produce similar results. However, because CL also estimates the relation (coefficient and statistical significance) between y_t^Q and x_t, it may produce a more accurate result, so CL is usually preferred. Nonetheless, because the DC model is widely used and intuitively attractive, we apply and test both

models. The BM of both CL and DC is computed using R (R Core Team, 2022), using the R package "tempdisagg" (Sax & Steiner, 2013) and the function "td". In this function we set "`method = chow-lin-maxlog`" for CL and "`method = denton-cholette`" for DC. The technical details of both BM models are further discussed in Sax and Steiner (2013).

5.2.3 Nowcasting models

In literature a wide variety of nowcasting models is discussed. An extensive literature overview is provided by the Handbook of Rapid Estimates (Eurostat, 2017). This Handbook also discusses mixed frequency models, which combine both temporal disaggregation/BM and nowcasting models by dealing with both issues simultaneously as a single missing data problem. For instance, the CL and DC BM models from the previous section can also be considered a mixed frequency NC model. Many of the advanced mixed frequency models are designed to deal with larger sets of auxiliary series and apply multivariate estimation. For example, recently Antolin-Diaz, Drechsel, and Petrella (2021) propose a Bayesian Dynamic Factor model that allows for time series with different frequencies to estimate daily GDP growth. Their model allows the use of a large set of time series and takes things like movements in long-run growth, time-varying uncertainty, and fat tails into account, by utilizing lag-lead properties of, and correlations between, auxiliary macroeconomic series with different frequencies. Frequentist versions of dynamic factor time series models are proposed by Giannone, Reichlin, and Small (2008) and Doz, Giannone, and Reichlin (2012). Another option is to use a vector autoregression (VAR) model, which estimates different PPC series simultaneously, (see e.g. Sims, 1980; Stock & Watson, 1980). These complex and data intensive models are beyond the scope of this paper, as we only consider the case where for each nowcast only one high-frequency auxiliary series and one low-frequency target series, which both measure the same phenomenon, are used. This simple approach has the advantage that the resulting estimates have a relatively straightforward interpretation, because the estimates do not depend on a large set of auxiliary series and no mutual dependencies between different PPC series are introduced.

We separate the NC models discussed in this paper into two groups. The first group we refer to as NC after BM models, which are the models that nowcast the high-frequency BM series directly with the help of the auxiliary series. We will discuss them in Section 5.2.3.1. The second group we refer to as *NC before BM* models. These models first nowcast the quarterly and monthly series for the current quarter, and then apply BM to obtain a nowcast for the current month. We will discuss them in Section 5.2.3.2.

5.2.3.1 NC after BM

The most basic nowcasting model we consider is a simple extrapolation (SE) nowcast model. This can be written as:

$$\hat{y}^{M,\text{BM,SE}}_{t|T_m} = \hat{y}^{M,\text{BM}}_{T_q|T_m}\left(x_t/x_{T_q}\right) \text{for } t = Q(T_m),\ldots,T_m. \tag{5.2}$$

This straightforward nowcasting approach is equivalent to the mixed frequency nowcasting model that results from extrapolating the DC BM model. However, equation (2) is slightly more general in the sense that $\hat{y}^{M,\text{BM}}_{T_q|T_m}$ can also be the result of a CL (or any other) BM model.

The second nowcasting model follows directly from the CL BM model, which can also be considered a mixed frequency NC model. CL performs linear regression on the quarterly level with y^Q_t and x^Q_t (x^Q_t is x_t aggregated to the quarterly level) and the estimated linear relation with x_t can be used to extrapolate over $t = Q(T_m),\ldots,T_m$. This can be written as:

$$\hat{y}^{M,\text{BM,CL}}_{t|T_m} = \beta^{\text{CL}}\mathbf{x}_t \text{ for } t = Q(T_m),\ldots,T_m, \tag{5.3}$$

with β^{CL} the CL regression coefficient.

A third type of nowcasting model is the well-known (seasonal) autoregressive-integrated moving average ((S)ARIMA) model (Box, 2013). (S)ARIMA can also incorporate auxiliary variables to obtain a nowcast of a target series. To select an appropriate (S)ARIMA model, a standardized stepwise procedure explained in Hyndman and Khandakar (2008) is used. This method is implemented in their R-package "`forecast`" and is used in this paper to obtain the nowcast $\hat{y}^{M,\text{BM,ARIMA}}_{t|T_m}$ and $\hat{y}^{M,\text{BM,SARIMA}}_{t|T_m}$. We apply both models because it is unclear whether the auxiliary variable can cover the seasonal pattern, which is present in most economic time series.

Another method which also applies SARIMA modeling is known as the Benchmark-to-Indicator-ratio (BIR) model (Bloem, Dippelsman, & Maehle, 2001; Daalmans, 2018). Its first step is to estimate a SARIMA model of the ratio $\hat{y}^{M,\text{BM}}_{t|T_m}/x_t$ for $t = 1,\ldots,T_q$, then obtain a SARIMA nowcast of the ratio series $\hat{y}^{M,\text{BM,BIR}}_{t|T_m}/x_t$ for $t = Q(T_m),\ldots,T_m$ and finally obtain a nowcast $\hat{y}^{M,\text{BM,BIR}}_{t|T_m}$ by multiplying by x_t for $t = Q(T_m),\ldots,T_m$. The BIR model might give better results when the ratio between $\hat{y}^{M,\text{BM,BIR}}_{t|T_m}$ and x_t is fixed but should be used with care when x_t can have values that are close to zero.

5.2.3.2 NC before BM

NC before BM models first perform a NC model on x_t and y^Q_t to obtain $\hat{x}_t$ and $\hat{y}^Q_t$ and then use these to obtain $\hat{y}^{M,\text{BM,NC}}_{t|T_m}$ with a BM model. This approach might be advantageous when the relation between the series y^Q_t and x_t is stronger on the quarterly level. To obtain $\hat{y}^Q_{t|T_m}$, we consider the Bridge, Mixed Data Sampling (MIDAS),

and Structural Time Series (STS) models. In the Bridge and MIDAS model, $\hat{x}_t$ is estimated with a univariate SARIMA model with x_t as input, while in the STS model it is estimated simultaneously with $\hat{y}_t^Q$.

The Bridge model, see Baffigi, Golinelli, and Parigi (2004); Angelini, Banbura, and Rünstler (2008) for extensive details, consists of a series of SARIMA models. First, $\hat{x}_t$ is estimated. Next, a SARIMA model is estimated with y_t^Q as dependent and x_t^Q as auxiliary variable. Then, by using $\hat{x}_t^Q$, an estimate of $\hat{y}_t^Q$ is obtained. Finally, a BM model with $\hat{x}_t$ and $\hat{y}_t^Q$ gives the estimate $\hat{y}_{t|T_m}^{M,\text{BM,Bridge}}$.

The MIDAS model approach (see Ghysels, Santa-Clara, & Valkanov, 2004; Ghysels, Sinko, & Valkanov, 2007, for extensive details) provides (just like the Bridge model) a quarterly estimate $\hat{y}_{t|T_m}^{Q,\text{MIDAS}}$ and a monthly estimate $\hat{y}_{T_m|T_m}^{\text{M,BM,MIDAS}}$. The MIDAS model is a regression and filtering technique that incorporates different frequencies. The difference with the Bridge model is that the MIDAS model allows for the modelling of lags in both the quarterly and monthly series simultaneously. According to the literature (see e.g. Asimakopoulos, Paredes, & Warmedinger, 2008), an advantage of the MIDAS model as compared to some alternatives such as state space and mixed frequency VAR models, is that the MIDAS model is more parsimonious and less sensitive to specification errors due to the use of non-linear lag polynomials. We estimated the MIDAS model in the R package "`midasr`", see Ghysels, Kvedaras, and Zemlys (2016) and the function "`midas_r`" with some basic settings.

A Structural Time Series (STS) model approach (see Durbin & Koopman, 2012), is not focused on obtaining $\hat{y}_{t|T_m}^{Q,\text{NC}}$ alone, but instead decomposes a time series into a trend, a seasonal component and additional noise. The details of the STS model for this application are described in the supplemental file. In this paragraph we only highlight the novel aspect of modelling the seasonal component. Both y_t^Q and x_t with $t = 1,\ldots,T_m$ are used as input of a multivariate STS model, i.e. both series are modelled jointly in a bivariate setting. Whereas for the other methods the quarterly value is repeated 3 times in the quarterly series, for the STM the value of this series is missing in the first and second month of each quarter. The quarterly value is used in the third month of each quarter. The STS model approach allows for missing values in the series. In this application, there are missing values in the last quarter, the model estimates for these periods are used as nowcasts. Similarly, as under the other approaches discussed, the auxiliary monthly series is used to improve the accuracy of the nowcasts. But under the STS model approach, this happens by explicitly modelling a correlation between trend disturbance terms of both series. We test two different trend models, the local and smooth trend model. For the local trend model, in the case of zero correlation between y_t^Q and x_t, the predictions are a flat line. For the smooth trend model, in the case of zero correlation between y_t^Q and x_t, the predictions are a linearly increasing or decreasing trend. In both models, in case of non-zero correlation, the predictions are adjusted by the auxiliary monthly series. The STS model can only disaggregate the trend component in monthly estimates. For the seasonal

component, an additional BM step, just like with the Bridge and MIDAS model, is required.

We developed a new approach to model the seasonal component of the target series y_t^Q. This seasonal model, which takes missing values into account, is an extension of the known dummy seasonal model.

$$s_t^y = \begin{cases} S_t^y & \text{if } t \text{ is the third month of the quarter} \\ 0 & \text{if } t \text{ is the first or second month of the quarter,} \end{cases} \tag{5.4}$$

with

$$S_t^y = \begin{cases} S_{t-10}^y & \text{if } t \text{ is the first month of the quarter} \\ S_{t-11}^y & \text{if } t \text{ is the second month of the quarter} \\ -S_{t-3}^y - S_{t-6}^y - S_{t-9}^y - \omega_t^y & \text{if } t \text{ is the third month of the quarter,} \end{cases} \tag{5.5}$$

$$E(\omega_t^y) = 0, \tag{5.6}$$

$$\text{Cov}(\omega_t^y, \omega_{t'}^y) = \begin{cases} \sigma_{\omega,y}^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases}$$

With (5.4) - (5.6) only the quarterly seasonal pattern of the target series y_t^Q can be estimated. The seasonal component is related with the observed series in the measurement equation of the state space model through s_t^y defined in (5.4). Since the observations are missing for the first two months of every quarter, s_t^y is equal to zero for the first two months and set equal to the quarterly pattern defined with s_t^y in the third month of the quarter. Equation (5.5) defines a quarterly seasonal pattern for the months. It is assumed that the monthly seasonal pattern is constant within each quarter. The first two rows of (5.5) show that during the first two months of a quarter, the seasonal pattern is equal to the value of the quarter in the previous year. The third row of equation (5.5) is like the standard dummy seasonal model for a quarterly series. In the third month of each quarter, except for the last quarter, the quarterly observation becomes available and the seasonal pattern for the last quarter (s_t^y) is updated using the values of the previous three quarters (S_{t-3}^y, S_{t-6}^y and S_{t-9}^y) and small change via ω_t^y. The seasonal pattern of the monthly auxiliary series x_t is modelled with a standard trigonometric seasonal component defined at a monthly frequency.

Under the assumption that the seasonal patterns of the monthly declarants and the quarterly declarants is similar, it is desirable that this monthly pattern is adopted by the quarterly series. Nevertheless, (smaller) differences between the seasonal patterns of the monthly and the quarterly series should be considered. This cannot be achieved with the structural time series model. Instead, DC or CL can be applied. It is expected that DC is suboptimal because this model cannot handle negative values easily.

The other components of the STS model are standard. Some adjustments are needed to take the relationship of the monthly and the quarterly figures of the y_t into account. Other adjustments are necessary to consider that the input series are partly based on the same enterprises.

The ideas behind the STS model approach are similar to the ideas behind the Bridge model approach. In both cases the auxiliary series is used to predict the target series. There are, however, some differences. First, in the STS model approach the auxiliary series is included as another dependent series and a correlation between trend disturbance terms of target series and auxiliary series is modelled. Both this correlation and the regression parameter in the Bridge approach are assumed to be constant over time. Second, in the STS approach all series are modelled and estimated simultaneously, including prediction of auxiliary series and target series and BM of the trend. Only the BM of the seasonal pattern is performed afterwards.

It might be worthwhile to express the various steps required to fit Bridge models in one state space model. The major advantage of such an approach is that it gives a more realistic approximation of the uncertainty of the nowcasts, since it avoids that estimates obtained in a particular step are treated as known in the next step. Casting a Bridge model in state space form requires that the BM of the target series in the final step is conducted with CL. Subsequently the target series and the auxiliary series are combined in a bivariate state space model, where both series are modelled with a SARIMA model, see Durbin and Koopman (2012, Ch. 3) for details. The SARIMA model for the target series must also include the auxiliary series as a regression component. At the same time the target series, observed at a quarterly frequency, must be modelled at a monthly frequency. Investigating the possibilities of this approach is left for further research.

5.2.4 Evaluation method

Altogether we can distinguish eleven different nowcasting models (i.e. SE, CL extrapolation, BIR, ARIMA, SARIMA, Bridge, MIDAS, and the local and smooth trend STS models with and without correlation) that are combined with two BM model variants. To compare their quality, a standard method to evaluate models is to make out-of-sample predictions and check how close these predictions are to the actual outcome. This can be measured by calculating e.g. the mean absolute error (MAE), which in this case can be computed at the monthly and quarterly level. We choose to look at the MAE because it is also applied in the quality assessment of the short-term business statistics at Statistics Netherlands.

As a benchmark series we use $\hat{y}^{M,\mathrm{CL}}_{t|T_{\max}}$, which is the with CL BM series of y^Q_t (with x_t) with $t = 1,\ldots,T_{\max}$, where $T_{\max}$ is simply the last month for which both monthly and quarterly data are available. This allows us to compare each $\hat{y}^{M,\mathrm{BM,NC}}_{T_m|T_m}$ with $\hat{y}^{M,\mathrm{CL}}_{T_m|T_{\max}}$, where $\hat{y}^{M,\mathrm{CL}}_{T_m|T_{\max}}$ is an estimate that is based on a maximum amount of available data.

The MAE of each series (denoted as $\text{MAE}^{M,\text{BM,NC}}$) can be written as:

$$\text{MAE}^{M,\text{BM,NC}} = \frac{\sum_{T_m=T_0}^{T_{\max}} \left| \hat{y}^{M,\text{BM,NC}}_{T_m|T_m} - \hat{y}^{M,\text{CL}}_{T_m|T_{\max}} \right|}{T_{\max} - T_0}, \tag{5.7}$$

where T_0 is the first month of the evaluation period. T_0 should not be too early in the time series, because each model requires a period of calibration. The quarterly MAE (denoted as $\text{MAE}^{Q,\text{BM,NC}}$) can be obtained by:

$$\text{MAE}^{Q,\text{BM,NC}} = \frac{\sum_{T_m=T_0}^{T_{\max}} \left| \left(\sum_{T_m=Q(T_m)}^{Q(t)+2} \hat{y}^{M,\text{BM,NC}}_{T_m|T_m} \right) /3 - \hat{y}^{Q}_{T_m} \right|}{3(T_{\max} - T_0)}, \tag{5.8}$$

The $\text{MAE}^{Q,\text{BM,NC}}$ has the advantage that the estimations are compared with the observed series y^Q_t instead of the estimated $\hat{y}^{M,\text{CL}}_{t|T_{\max}}$.

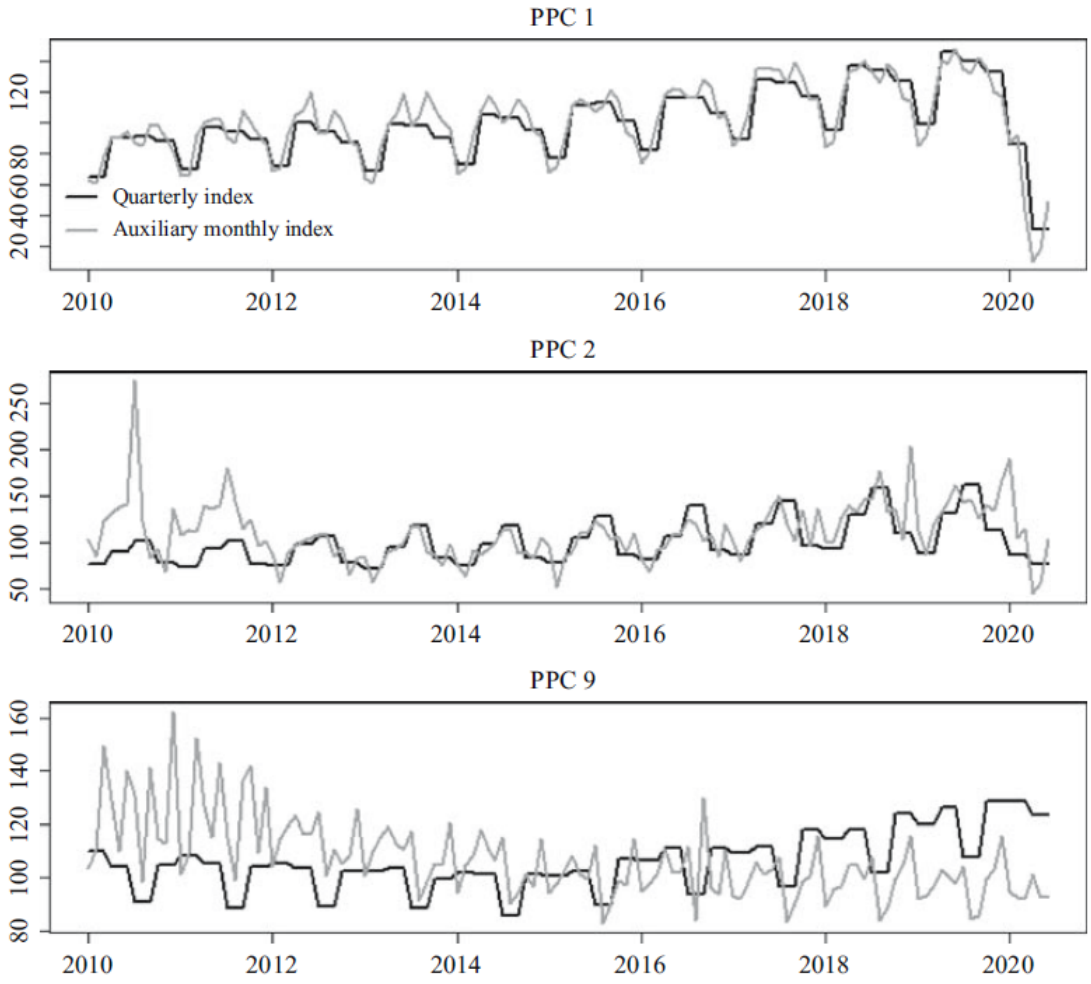
5.3 Empirical evaluation of the nowcast models

In this section we apply and compare the models that were introduced in Section 5.2. We first describe the data in Section 5.3.1. Next, in Section 5.2.4, we discuss how the models compare in terms of their MAE. In the last subsections (5.3.2 and 5.3.3) we discuss how they perform before, during and after a simulated economic shock.

5.3.1 Time series data

We apply all models from Section 5.2 on index time series from twelve PPCs, that cover the period January 2010 until June 2020, and that represent twelve different economic activities in the Netherlands based on four-digit NACE (Nomenclature statistique des activités économiques dans la Communauté européenne) (Eurostat, 2008). Six of them represent the hospitality sector (i.e. Hotels, Other accommodation, Restaurants, Fast food, Catering and Pubs) and six of them represent other activities in the service sector (i.e. Publishers, Legal activities, Accountants, Employment activities, Other Business Support, Repair of household goods). We refer to them as PPCs 1 – 12, in the same order as above. For all twelve series both y^Q_t and x_t are available. The auxiliary monthly index series x_t is based on the raw turnover data from monthly declarants that declared turnover in all consecutive months of the series and is only corrected for new and bankrupt companies. The published quarterly index series y^Q_t can be considered of higher quality, because it is based on turnover data from both monthly and quarterly declarants, it is manually corrected for errors, for new and bankrupt companies, and is complemented with primary data collection for a small group of exceptionally large companies. Both y^Q_t and x_t are rescaled such that they have a mean value of 100 in the year 2015.

Figure 5.1: Published quarterly series and monthly auxiliary index series of PPC 1, 2 and 9, over the period January 2010 – June 2020



The series y_t^Q and x_t for PPC 1, 2 and 9 are presented in Figure 5.1 below. These PPCs illustrate how similar or different both series can be. In the graphs of PPC 1, 2 and 9 we see that y_t^Q and x_t can be correlated to different degrees. The graphs also show that the seasonal pattern generally presents itself in both the monthly and the quarterly series. However, we also see that the relation between the series may or may not be stable over the entire period.

Furthermore, we see that the different PPCs are affected differently by the COVID-19 pandemic in the second quarter of 2020. For example, Hotels (PPC 1) show a big collapse in both the monthly and the quarterly series, while Accountants (PPC 9) seem to be hardly affected. The COVID-19 pandemic raises the question whether the models that perform best during a stable economic period also perform best during a crisis. This is investigated by evaluating the nowcast models separately over the period July 2016 until February 2020 (stable period) and over the period March 2020

until June 2020 (crisis period).

5.3.2 Nowcast model performance before and during a crisis

Table 5.1 shows the nowcasting results for all BM and NC model combinations and PPCs. The last two columns show the unweighted and weighted (by annual turnover in 2015) results over all PPCs.

Table 5.1: $\text{MAE}^{M,\text{BM},\text{NC}}$ over July 2016 – February 2020 for all 12 PPCs, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	6.5	15.2	4.7	6.1	8.6	4.9	7.1	7.1	7.0	3.5	4.7	2.8	6.5	5.5
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.4	3.3	0.5	0.7	0.6	1.0	1.2	0.3	1.2	3.0	2.7	2.1	1.4	1.7
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	2.3	26.0	1.8	8.6	4.3	4.2	11.8	4.8	9.1	4.2	5.9	6.6	7.5	5.9
			CL extrapolation	3.2	22.0	1.7	7.5	4.4	5.2	9.9	5.5	10.4	4.7	7.9	5.3	7.3	6.3
			ARIMA	2.9	20.0	1.6	7.1	4.6	3.9	8.5	4.8	8.2	4.2	5.4	4.9	6.4	5.3
			SARIMA	3.1	18.4	1.5	6.8	3.6	4.0	7.8	4.1	6.7	4.3	4.9	4.2	5.8	4.9
			BIR	3.1	13.1	1.5	5.5	4.3	3.1	4.5	2.4	5.6	6.9	5.6	7.5	5.3	5.2
		DC	Simple extrapolation	2.1	25.7	1.7	9.5	4.3	3.1	12.9	4.8	9.2	4.9	6.6	7.1	7.7	6.3
			ARIMA	1.9	20.5	1.4	7.2	3.7	3.3	7.9	5.8	6.9	4.7	3.9	7.3	6.2	5.2
			SARIMA	1.9	10.9	1.2	5.9	4.1	2.5	4.6	2.4	4.8	4.9	4.5	5.6	4.4	4.2
			BIR	2.1	22.0	1.5	8.7	4.1	2.9	10.2	3.7	6.0	4.9	5.4	6.4	6.5	5.3
	NC before BM	CL	Bridge	3.1	6.1	1.3	3.3	3.2	2.2	3.6	1.8	3.8	4.0	3.2	2.9	3.2	3.3
			MIDAS	3.2	6.8	1.4	3.5	3.5	2.1	4.4	1.6	3.8	5.0	3.8	3.6	3.5	3.8
		DC	Bridge	3.1	5.6	1.3	5.0	3.0	2.3	3.9	1.9	3.9	4.4	4.1	2.6	3.4	3.6
			MIDAS	3.2	6.5	1.3	5.2	3.2	2.3	4.5	1.8	4.2	5.6	4.7	3.4	3.8	4.2
State-space		CL	STS, local trend model, no correlation	3.0	13.5	2.6	4.7	3.7	3.9	5.5	3.7	6.6	3.1	4.2	3.6	4.8	4.2
			STS, smooth trend model, no correlation	2.0	13.2	2.1	4.6	3.6	4.0	5.7	3.4	6.3	2.6	4.3	3.4	4.6	3.8
			STS, local trend model, correlation	2.6	13.6	1.8	4.7	3.6	3.9	5.6	3.7	6.6	2.6	4.1	3.4	4.7	3.9
			STS, smooth trend model, correlation	1.9	13.0	1.8	4.7	3.5	4.0	5.8	3.4	6.2	2.6	4.2	3.4	4.5	3.8
		DC	STS, local trend model, no correlation	4.2	25.9	3.6	10.4	3.7	5.0	6.4	8.3	11.1	3.4	6.3	8.0	8.0	6.2
			STS, smooth trend model, no correlation	2.9	26.4	3.0	10.4	3.5	5.2	6.8	8.0	10.9	4.3	6.7	7.8	8.0	6.4
			STS, local trend model, correlation	4.2	26.4	2.6	10.3	3.5	4.5	6.7	8.3	11.1	3.2	6.2	7.9	7.9	6.0
			STS, smooth trend model, correlation	3.4	25.9	2.8	10.5	3.3	5.4	7.1	8.0	10.6	4.3	6.1	7.8	7.9	6.3

In Table 5.1 the various models are evaluated on the monthly level during the stable economic period July 2016 – February 2020. The grey cells in Table 5.1 are the cells in the top three of lowest MAEs in each PPC column. The rows that represent the Bridge and MIDAS models contain most of these grey cells, which implies they generally perform better than the other models in the table. The best model according to both the unweighted and weighted mean, is the Bridge model combined with the Chow-Lin BM model, while the STS models in combination with DC perform clearly worse.

Table 5.2 shows MAEs for each model/PPC combination, but now the MAE is calculated over the quarters in the period July 2016 – December 2019 (The last quarter that was unaffected by the COVID-19 pandemic).

Table 5.2 yields the same conclusion as Table 5.1, as again the Bridge and MIDAS model perform quite well. To illustrate this graphically, Figure 5.2 shows $\hat{y}^{M,\text{CL}}_{t|T_{\max}}$ and $\hat{y}^{M,\text{CL},\text{Bridge}}_{T_m|T_m}$ for PPC 1, 2 and 9. It shows that $\hat{y}^{M,\text{CL},\text{Bridge}}_{T_m|T_m}$ performs quite well for all

Table 5.2: $MAE^{Q,BM,NC}$ over July 2016 – December 2019 for all 12 PPCs, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	2.1	26.0	1.8	8.1	4.1	4.1	11.3	4.9	8.8	2.6	5.1	6.1	7.1	5.2
			CL extrapolation	3.0	21.3	1.7	6.9	4.0	5.3	9.2	5.5	10.0	4.7	7.5	5.2	7.0	6.1
			ARIMA	2.7	19.1	1.6	6.5	4.3	4.0	8.1	4.8	8.0	4.2	4.9	4.8	6.1	5.2
			SARIMA	2.8	17.6	1.5	6.4	3.2	4.1	7.2	4.2	6.5	4.3	4.4	4.1	5.5	4.8
			BIR	2.8	11.5	1.4	5.2	4.1	3.1	4.0	2.5	5.4	7.1	4.6	7.4	4.9	5.1
		DC	Simple extrapolation	1.8	26.2	1.7	9.0	4.0	2.9	12.2	4.9	9.0	3.3	5.7	7.0	7.3	5.6
			ARIMA	1.8	20.5	1.4	6.2	3.4	3.2	7.4	6.0	6.6	3.0	3.6	7.3	5.9	4.5
			SARIMA	1.6	10.0	1.1	5.3	3.8	2.4	3.9	2.4	4.1	3.5	4.2	5.4	4.0	3.4
			BIR	1.8	22.2	1.5	8.3	3.8	2.8	9.8	3.8	5.2	3.0	4.6	6.2	6.1	4.5
	NC before BM	CL	Bridge	3.0	3.3	1.3	2.6	2.9	1.9	2.3	1.2	3.0	3.8	2.4	2.3	2.5	2.8
			MIDAS	3.0	3.5	1.3	2.6	2.9	1.9	2.2	1.2	3.1	4.0	2.3	2.5	2.6	2.9
		DC	Bridge	3.0	3.4	1.0	4.5	2.7	2.3	2.6	1.4	3.7	4.3	3.7	2.1	2.9	3.3
			MIDAS	3.0	3.2	1.0	4.5	2.7	2.2	2.5	1.5	3.9	4.7	3.7	2.3	2.9	3.4
State-space		CL	STS, local trend model, no correlation	2.8	12.0	2.6	3.6	1.5	3.5	2.1	2.8	5.0	2.8	3.6	2.7	3.7	3.3
			STS, smooth trend model, no correlation	1.9	11.7	1.7	3.7	1.7	3.6	2.2	2.5	4.4	2.0	3.2	2.5	3.4	2.8
			STS, local trend model, correlation	2.4	12.0	1.7	3.7	1.4	3.4	2.2	2.8	5.0	2.2	3.5	2.5	3.6	3.0
			STS, smooth trend model, correlation	1.7	11.4	1.6	3.7	1.5	3.6	2.3	2.5	4.3	2.0	3.1	2.5	3.3	2.7
		DC	STS, local trend model, no correlation	3.2	17.5	3.4	7.1	1.9	4.4	5.0	6.0	5.8	3.0	5.4	4.3	5.6	4.5
			STS, smooth trend model, no correlation	1.8	19.7	2.2	7.3	1.6	4.9	4.8	6.1	6.0	2.7	5.7	4.0	5.6	4.3
			STS, local trend model, correlation	3.2	17.6	2.3	7.1	1.7	4.0	5.1	6.0	5.9	2.4	5.5	4.2	5.4	4.2
			STS, smooth trend model, correlation	2.0	19.5	2.2	7.3	1.5	5.1	5.1	6.1	6.0	2.7	5.1	3.9	5.5	4.3

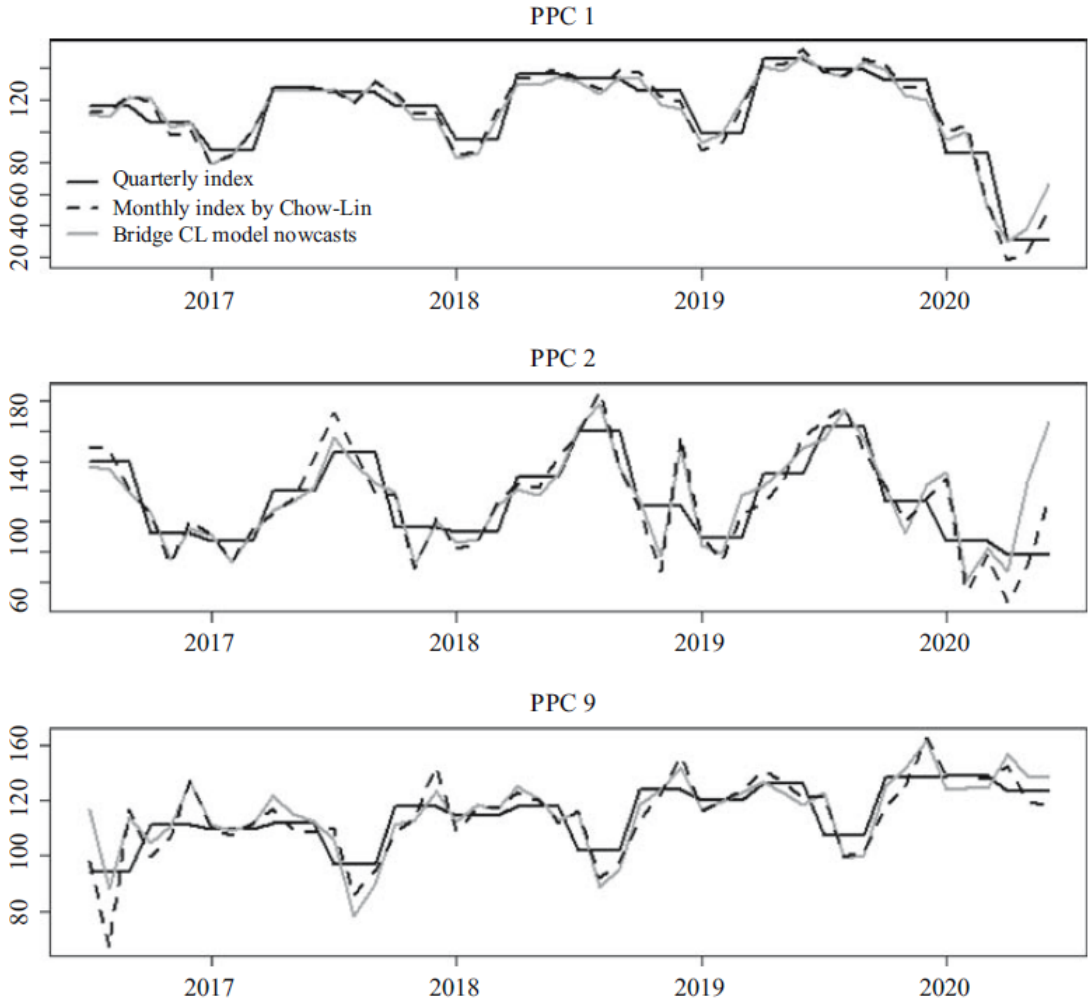
three PPCs, except it underestimates the impact of the COVID-19 pandemic in the last few months.

To investigate whether a different nowcast model should be preferred during a crisis, Table 5.3 shows the MAEs for the period March 2020 – June 2020. Table 5.3 shows that during a crisis, the Bridge and MIDAS model that were fitted on data that is largely from periods prior to the crisis, no longer provide the most accurate nowcasts, but the more basic models perform somewhat better. This is not surprising, because they rely less on the past and more on recent data. The DC, (S)ARIMA method is the most accurate method among the direct ones. Surprisingly, the DC, (S)ARIMA model is clearly more accurate than the CL, (S)ARIMA model. The explanation lies in the AR(1) term that is part of both the CL and (S)ARIMA model. Therefore, the CL, (S)ARIMA model puts more weight on the past than the DC, (S)ARIMA model. This is a disadvantage during a crisis.

To further investigate how robust the above results are with respect to the MAE evaluation method, two other evaluation methods are applied. The first method counts in how many periods a specific method is more accurate than CL, Bridge (before the COVID-pandemic) or DC, SARIMA (during the first months of the COVID-pandemic). The second measure counts how often the relative prediction error is over 8%. These alternative evaluation methods confirm that the earlier results based on the MAE. Details about these alternative evaluation methods and the results are available from the authors on request.

A question that remains unanswered in this section, is which model should be

Figure 5.2: Quarterly, monthly and Bridge CL NC index series of PPC 1, 2 and 9, over the period July 2016 – June 2020.



preferred after a crisis? How long after a crisis will the CL, Bridge model start to outperform the DC, SARIMA model again? This question is the subject of the next section.

5.3.3 Nowcast model performance after a crisis

To investigate the performance of the models during a longer crisis and after a crisis, we simulate three different types of economic shocks in January 2017. Each shock implies that both $y^{Q}_{(\text{January 2017})}$ and $x_{(\text{January 2017})}$ are divided by 2. The first type (type 1) does not recover, the second (type 2) recovers during the next quarter and the third (type 3) slowly recovers during a period of one year. Figure 5.3 shows these three shocks with PPC 1 as an example. In this simulation study the COVID-19 pandemic crisis period is excluded from the analysis.

Table 5.3: $MAE^{M,BM,NC}$ over March 2020 – June 2020 for all 12 PPCs, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	18.2	20.6	23.3	5.3	12.1	22.6	3.5	3.2	4.6	3.2	5.3	4.8	10.6	8.0
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		4.6	8.1	10.0	6.2	1.1	11.8	4.4	0.5	0.9	3.0	3.1	5.8	5.0	4.0
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	7.3	28.3	10.0	30.0	7.8	13.6	13.4	7.9	12.1	2.0	15.3	7.5	12.9	9.2
			CL extrapolation	6.9	27.6	6.9	28.2	6.2	5.6	22.5	7.8	7.0	4.9	7.8	6.6	11.5	8.9
			ARIMA	6.5	18.0	6.4	29.2	6.6	7.0	18.7	6.2	2.9	5.6	6.4	4.7	9.9	7.9
			SARIMA	6.7	23.9	6.4	29.0	6.2	7.8	15.5	5.7	2.7	5.5	6.1	4.7	10.0	7.7
			BIR	11.1	28.6	9.2	26.5	6.5	15.5	9.9	2.2	2.9	1.7	9.5	8.5	11.0	6.9
		DC	Simple extrapolation	5.9	26.6	12.0	30.4	7.2	10.9	11.6	7.5	13.0	5.4	12.4	6.2	12.4	10.0
			ARIMA	2.2	23.3	12.6	28.3	4.8	4.2	7.1	6.7	4.2	2.7	2.9	3.8	8.6	6.2
			SARIMA	2.5	26.0	12.5	26.9	4.4	3.4	5.3	1.4	3.8	2.4	2.2	4.9	8.0	5.4
			BIR	5.6	23.2	12.5	31.1	6.4	9.8	10.6	6.0	5.9	2.7	8.1	6.6	10.7	7.7
	NC before BM	CL	Bridge	11.4	32.6	3.7	6.7	23.4	43.4	6.6	2.9	6.7	7.5	6.1	4.7	13.0	9.3
			MIDAS	14.4	34.6	7.6	10.7	22.3	58.7	7.3	2.6	6.0	8.2	6.2	8.4	15.6	10.9
		DC	Bridge	71.7	31.4	16.1	17.3	27.0	15.4	4.4	2.9	7.7	14.0	9.6	8.6	18.9	16.0
			MIDAS	76.6	32.4	11.3	19.1	25.8	26.7	6.9	2.6	7.2	14.6	8.8	10.9	20.2	16.7
State-space		CL	STS, local trend model, no correlation	86.0	21.3	61.1	18.6	49.9	47.2	5.2	2.9	6.3	17.1	10.6	5.1	27.6	24.0
			STS, smooth trend model, no correlation	93.2	24.9	73.3	24.8	49.9	47.3	4.9	2.5	5.8	17.8	13.5	7.4	30.4	26.3
			STS, local trend model, correlation	37.7	18.8	2.0	12.9	34.1	36.1	3.5	2.7	6.3	12.8	9.8	4.8	15.1	12.6
			STS, smooth trend model, correlation	52.9	27.7	21.8	6.6	37.8	40.9	4.0	2.5	5.7	15.9	12.7	7.3	19.7	16.8
		DC	STS, local trend model, no correlation	90.3	31.8	68.7	22.1	52.2	41.5	6.0	4.9	7.0	17.6	9.5	5.9	29.8	25.6
			STS, smooth trend model, no correlation	112.8	34.3	127.3	25.0	53.8	42.5	4.8	3.8	6.8	18.3	12.2	8.7	37.5	33.0
			STS, local trend model, correlation	39.6	32.1	2.4	32.0	35.6	49.9	6.4	4.7	7.0	13.0	8.7	5.1	19.7	15.0
			STS, smooth trend model, correlation	70.9	27.6	31.5	15.6	40.6	51.8	5.1	3.9	8.3	16.3	11.9	8.6	24.3	20.4

To investigate the direct and long-term effect of a shock on the performance of different models, we calculate the MAE separately over the period January 2017 – December 2017 for shock 1 and 3 (results in Table 5.4 and 5.7), over the period January 2017 – March 2017 for shock 2 (results in Table 5.5) and January 2018 – December 2019 for shock 3 (results in Table 5.7). Just like during the COVID-19 pandemic crisis, the NC after BM models outperform the NC before BM models.

Table 5.4 - 5.6 show that the BIR model performs among the best three models (according to both the weighted and unweighted mean) during all three shocks. This can be partly explained by the simulation setup because the BIR model is based on predicting the ratio $\hat{y}_{t|T_m}^{M,\mathrm{BM}}/x_t$, which is by construction hardly disturbed by our artificial shocks, because both y_t^Q and x_t are multiplied with the same factor. In a real crisis, both series might be affected in different ways, which could make the other methods more competitive, as was seen in Table 5.3. Furthermore, Table 5.4 - 5.6 show that some models have a serious problem in nowcasting the second type of shock, leading to very high MAEs. This concerns all the NC before BM models, the CL extrapolation and CL, (S)ARIMA models, which show large mean MAEs due to large MAEs for PPC 7, 10 and 12. The main reason for this last result is that the CL, (S)ARIMA model estimates a correlation between y_t^Q and x_t, which might be overestimated due to the artificial shock in both series. A final point of interest is that the NC after BM models have more problems with the gradual recovering shock 3 than shock 1 and 2.

Next, in Table 5.7 we look at the MAEs of the different models in the second and third year after shock 3. Just as in the analysis of the real data in Section 5.3.2, the

Figure 5.3: Illustration of three types of shocks (as of January 2017) with PPC 1 as example.

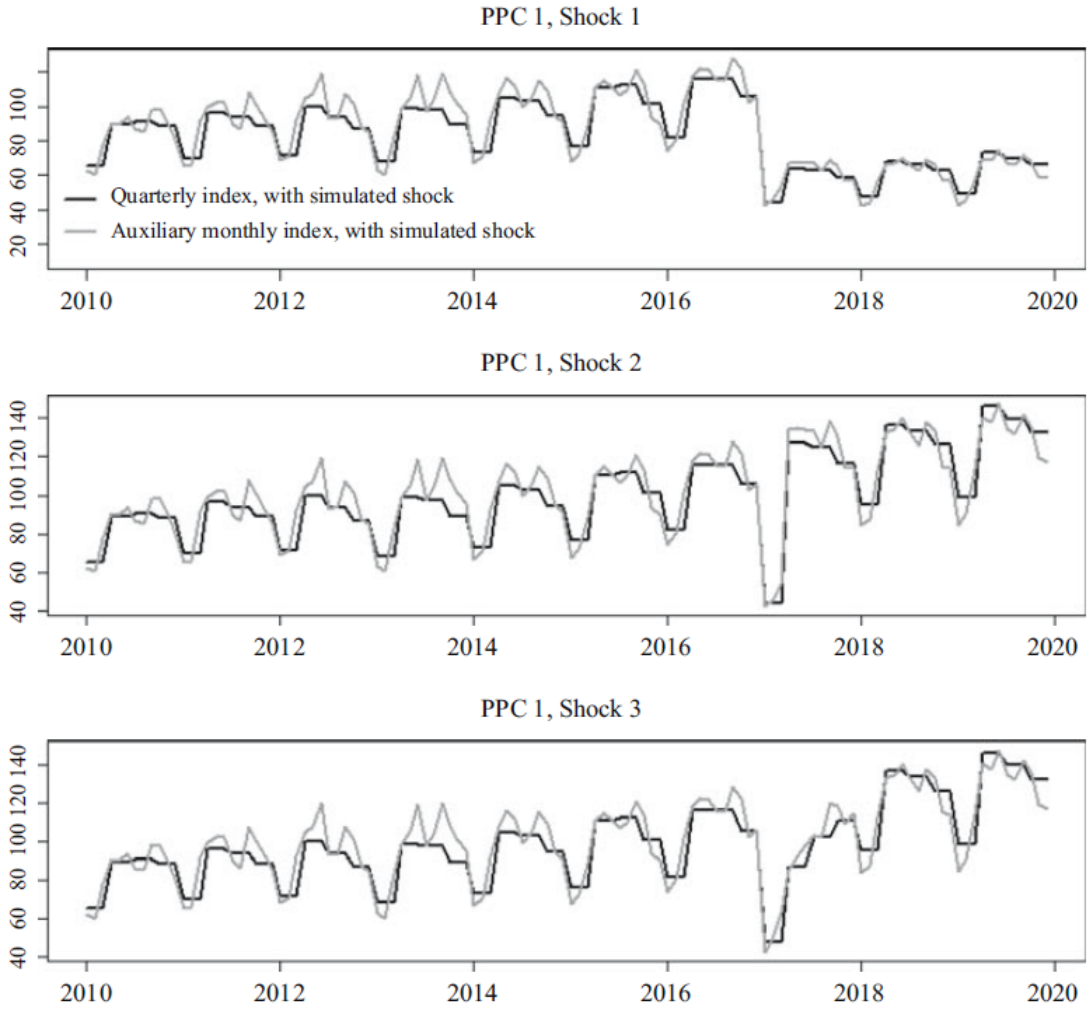


Table 5.4: $MAE^{M,BM,NC}$ over January 2017 – December 2017 for all 12 PPCs with shock 1, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	2.4	7.2	2.2	3.1	3.6	1.8	3.7	3.8	3.0	1.7	3.9	2.6	3.2	2.7
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.1	1.5	0.1	0.2	0.4	0.4	0.7	0.2	1.0	2.0	0.6	0.6	0.7	1.0
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	0.8	12.6	0.7	3.6	3.2	2.2	5.6	3.3	2.7	3.8	3.2	5.0	3.9	3.4
			CL extrapolation	0.7	11.0	1.1	3.2	4.0	2.7	8.6	4.2	4.1	12.3	4.1	11.3	5.6	6.9
			ARIMA	0.7	8.2	1.0	2.6	3.8	2.5	8.8	3.8	3.3	12.2	2.2	10.5	5.0	6.5
			SARIMA	1.0	7.6	0.9	3.0	3.1	2.0	8.2	4.1	3.3	11.5	2.4	10.6	4.8	6.2
			BIR	0.7	7.3	0.6	2.4	3.3	1.6	2.0	2.7	2.9	5.5	3.2	4.8	3.1	3.4
		DC	Simple extrapolation	0.9	11.8	0.7	3.8	2.9	1.6	5.6	3.3	3.3	2.6	3.7	4.5	3.7	3.0
			ARIMA	0.8	11.9	1.0	2.4	2.2	5.0	6.6	5.9	9.9	5.7	6.3	8.3	5.5	5.5
			SARIMA	0.8	5.0	0.5	2.5	2.2	4.9	3.8	3.2	11.8	4.3	6.3	8.3	4.5	4.6
			BIR	0.7	10.3	0.7	3.7	2.5	1.6	4.7	3.2	1.5	2.8	2.8	4.0	3.2	2.7
	NC before BM	CL	Bridge	1.3	32.2	5.5	11.2	13.3	8.6	8.1	7.0	15.6	12.1	12.9	17.1	12.1	11.0
			MIDAS	1.3	28.7	5.0	12.2	12.7	7.3	8.4	6.6	14.4	10.4	12.0	15.6	11.2	10.0
		DC	Bridge	5.0	33.0	8.0	7.9	10.4	8.4	11.3	7.1	12.0	13.5	11.9	19.3	12.3	11.4
			MIDAS	5.2	29.8	7.7	8.8	10.3	6.2	11.7	6.2	10.9	11.9	10.9	17.6	11.4	10.5
State-space		CL	STS, local trend model, no correlation	18.4	21.3	16.1	18.1	17.3	21.8	18.0	19.1	23.1	21.5	25.6	17.5	19.8	20.5
			STS, smooth trend model, no correlation	24.2	27.7	21.5	22.8	16.8	14.8	17.8	16.2	20.6	25.1	20.0	19.9	20.6	21.7
			STS, local trend model, correlation	6.4	20.3	4.0	8.1	11.1	9.8	10.5	9.5	14.9	15.5	17.2	15.2	11.9	12.5
			STS, smooth trend model, correlation	9.1	26.5	9.1	11.4	14.8	17.5	16.7	11.9	16.8	15.5	19.5	15.1	15.3	14.9
		DC	STS, local trend model, no correlation	25.2	38.6	14.9	30.5	18.2	29.8	17.1	19.3	27.6	24.6	24.2	16.6	23.9	23.6
			STS, smooth trend model, no correlation	29.7	43.2	25.7	24.2	17.7	25.0	19.1	17.7	29.0	25.7	19.1	17.9	24.5	24.7
			STS, local trend model, correlation	12.0	40.7	6.0	18.1	12.1	18.5	12.5	12.4	16.8	17.2	13.7	15.2	16.3	15.2
			STS, smooth trend model, correlation	13.2	45.3	12.7	15.0	15.8	27.1	18.1	15.2	23.0	16.7	19.0	16.3	19.8	18.0

Table 5.5: $MAE^{M,BM,NC}$ over January 2017 – March 2017 for all 12 PPCs with shock 2, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	3.9	4.1	3.7	3.7	1.2	2.9	4.5	6.4	1.4	5.3	5.4	4.7	3.9	4.2
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.1	0.9	0.1	0.9	0.5	0.6	1.3	1.0	1.0	2.7	0.9	0.8	0.9	1.4
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	0.8	5.4	0.3	2.2	2.4	2.8	8.9	1.7	1.7	2.5	3.3	1.3	2.8	2.6
			CL extrapolation	0.8	6.9	1.8	2.6	2.2	1.8	22.6	4.5	2.5	38.3	3.8	29.8	9.8	16.6
			ARIMA	1.0	0.6	1.9	2.4	1.8	4.9	24.4	4.7	5.4	37.6	0.7	29.4	9.6	16.6
			SARIMA	0.2	0.7	1.5	3.5	2.3	2.0	22.7	6.3	2.1	38.1	0.6	29.8	9.2	16.2
			BIR	0.6	2.8	0.2	1.5	0.7	1.9	3.4	1.7	5.0	5.1	3.9	2.9	2.5	3.3
		DC	Simple extrapolation	0.8	5.8	0.4	2.2	2.7	2.2	7.7	1.8	1.7	2.2	2.4	1.6	2.6	2.4
			ARIMA	0.8	17.2	1.7	2.5	0.9	14.1	6.9	10.8	17.5	9.2	12.1	14.3	9.0	9.0
			SARIMA	0.3	8.4	0.2	4.5	1.4	12.6	2.4	9.7	17.5	8.1	14.5	15.9	8.0	8.1
			BIR	0.7	3.3	0.4	1.8	1.7	1.9	6.8	2.0	1.1	2.2	2.6	1.1	2.1	2.1
	NC before BM	CL	Bridge	1.0	33.8	19.6	42.1	28.7	30.0	21.9	21.9	45.1	38.3	33.3	45.1	30.1	31.7
			MIDAS	0.7	29.6	17.8	46.3	26.4	25.0	20.1	19.9	38.6	32.6	31.0	39.7	27.3	28.1
		DC	Bridge	16.4	36.9	28.0	26.9	24.7	31.0	23.1	23.3	35.2	43.7	36.6	40.8	30.5	33.7
			MIDAS	17.1	32.4	26.6	29.4	23.0	21.8	22.1	20.1	29.7	38.5	33.7	35.7	27.5	30.1
State-space		CL	STS, local trend model, no correlation	33.9	31.1	35.6	41.7	42.3	33.9	42.2	43.1	43.5	46.7	48.6	44.5	40.6	42.7
			STS, smooth trend model, no correlation	35.9	34.1	42.0	48.7	42.8	36.9	41.0	42.9	46.9	47.4	52.3	44.0	42.9	44.7
			STS, local trend model, correlation	20.4	31.9	13.8	28.3	34.7	33.6	34.5	32.6	45.0	46.0	44.6	43.0	34.0	37.2
			STS, smooth trend model, correlation	27.3	33.9	31.7	38.0	39.0	35.8	35.6	36.4	43.7	48.5	52.0	41.4	38.6	41.6
		DC	STS, local trend model, no correlation	59.0	118.6	31.5	75.2	44.9	69.0	44.8	44.5	34.7	46.0	24.5	38.1	52.6	46.2
			STS, smooth trend model, no correlation	58.3	117.1	58.6	50.5	46.1	75.0	44.0	45.5	41.0	45.5	41.3	38.8	55.1	49.6
			STS, local trend model, correlation	42.3	118.9	21.3	53.9	37.1	68.7	36.2	33.8	36.5	46.8	24.4	36.3	46.4	41.9
			STS, smooth trend model, correlation	43.5	116.8	44.6	39.1	41.8	74.7	37.4	37.9	39.2	46.9	39.7	35.2	49.7	45.7

Table 5.6: $MAE^{M,BM,NC}$ over January 2017 – December 2017 for all 12 PPCs with shock 3, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	5.5	12.2	5.4	5.3	8.8	5.1	8.7	8.5	5.9	5.7	8.7	5.5	7.1	6.6
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.3	2.2	0.1	0.4	0.7	0.5	1.3	0.6	1.2	2.4	0.8	1.3	1.0	1.3
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	1.5	21.9	1.4	5.0	4.9	3.3	8.3	5.1	4.0	4.9	4.9	8.2	6.1	4.9
			CL extrapolation	1.3	17.9	1.6	4.8	6.5	4.8	10.4	5.7	7.2	15.3	6.8	13.1	7.9	9.2
			ARIMA	1.2	14.9	1.5	3.8	6.3	4.5	10.6	5.3	5.6	14.7	4.0	12.2	7.0	8.5
			SARIMA	1.5	13.5	1.4	4.1	5.5	4.2	9.4	5.3	3.4	15.1	4.5	12.2	6.7	8.2
			BIR	1.3	12.4	1.1	3.1	5.9	2.7	2.6	3.6	4.0	7.6	4.9	7.8	4.8	4.9
		DC	Simple extrapolation	1.8	20.9	1.3	5.4	4.1	2.5	8.9	5.3	4.7	3.2	5.5	7.8	6.0	4.5
			ARIMA	1.6	17.4	1.5	3.2	3.5	6.4	6.5	7.9	11.1	5.6	7.5	9.8	6.8	6.2
			SARIMA	1.5	7.3	0.9	3.6	3.7	4.4	2.9	4.8	9.8	4.3	8.6	9.7	5.1	4.8
			BIR	1.5	18.6	1.2	5.1	4.3	2.5	7.1	5.3	2.0	3.5	4.3	6.9	5.2	3.9
	NC before BM	CL	Bridge	1.7	18.4	5.6	10.6	11.7	11.5	9.0	7.0	13.3	12.4	13.5	15.1	10.8	10.6
			MIDAS	1.8	18.4	5.4	12.3	11.3	11.1	8.7	6.9	12.8	11.5	13.5	14.6	10.7	10.3
		DC	Bridge	5.9	24.4	8.1	9.9	10.2	9.0	11.5	8.9	10.8	12.3	12.4	15.4	11.6	11.0
			MIDAS	6.2	23.5	8.2	10.7	10.1	7.1	11.3	8.5	9.9	11.9	12.0	14.7	11.2	10.6
State-space		CL	STS, local trend model, no correlation	18.3	19.1	17.5	18.1	19.8	14.2	17.7	17.2	20.1	22.5	21.2	18.3	18.7	19.9
			STS, smooth trend model, no correlation	24.7	24.0	18.8	20.6	22.5	16.0	19.9	20.0	19.9	27.6	21.8	27.1	21.9	23.0
			STS, local trend model, correlation	6.5	19.3	4.1	8.7	10.9	12.6	11.9	10.9	16.1	14.3	18.4	12.9	12.2	12.6
			STS, smooth trend model, correlation	10.4	25.5	6.4	11.5	13.5	17.4	18.1	13.0	18.4	18.1	23.6	17.0	16.1	16.2
		DC	STS, local trend model, no correlation	24.2	49.5	16.6	28.9	20.6	23.4	18.0	18.4	19.9	24.4	16.9	18.0	23.2	22.1
			STS, smooth trend model, no correlation	29.2	55.1	23.2	23.0	22.2	25.7	23.9	22.1	21.4	30.2	22.1	22.4	26.7	26.4
			STS, local trend model, correlation	12.0	50.2	6.1	18.9	11.3	20.8	13.3	14.3	17.3	15.9	14.3	15.5	17.5	15.5
			STS, smooth trend model, correlation	14.0	57.6	6.6	15.3	14.4	26.9	23.6	17.3	22.1	18.7	24.3	16.8	21.5	19.2

Table 5.7: $MAE^{M,BM,NC}$ over January 2018 – December 2019 for all 12 PPCs after shock 3, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	5.9	15.3	4.2	5.1	9.1	4.7	6.9	6.5	6.4	4.5	6.4	4.5	6.6	5.7
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.3	3.1	0.2	0.4	0.5	0.7	1.2	0.3	1.1	1.1	0.9	0.5	0.8	0.8
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	2.5	26.7	2.0	10.3	3.9	3.7	10.4	5.5	6.6	4.1	4.4	7.9	7.3	5.5
			CL extrapolation	3.7	21.7	2.0	8.9	4.2	5.6	9.8	5.6	9.3	5.1	8.1	8.0	7.7	6.4
			ARIMA	3.4	19.0	1.9	8.2	4.8	3.9	8.5	4.9	6.7	4.6	5.4	6.8	6.5	5.4
			SARIMA	3.5	18.1	1.8	8.1	3.7	4.5	7.9	4.1	4.8	4.3	4.7	6.9	6.0	4.9
			BIR	3.4	11.6	1.6	5.9	5.0	3.0	4.1	2.6	3.4	4.0	4.7	7.1	4.7	3.9
		DC	Simple extrapolation	2.1	27.3	1.9	10.9	4.1	2.6	11.6	5.6	6.8	3.9	5.4	7.6	7.5	5.7
			ARIMA	2.0	20.0	1.5	7.6	4.5	4.3	7.8	6.4	4.6	3.8	3.8	7.5	6.1	4.7
			SARIMA	2.0	15.0	1.4	6.4	3.9	3.6	4.5	2.5	5.7	3.5	4.5	7.2	5.0	4.0
			BIR	2.2	22.9	1.7	9.8	3.9	2.6	8.8	4.1	3.6	3.9	4.4	6.9	6.2	4.7
	NC before BM	CL	Bridge	3.2	8.4	1.5	4.0	3.9	3.4	4.0	2.1	5.0	3.9	4.3	4.9	4.0	3.7
			MIDAS	3.2	9.9	1.5	4.1	3.8	3.3	5.0	2.1	4.7	4.0	4.2	4.9	4.2	3.9
		DC	Bridge	3.6	9.3	1.8	5.3	4.6	3.2	8.4	3.4	3.5	3.3	5.2	7.7	4.9	4.0
			MIDAS	3.5	9.9	1.7	5.4	4.6	3.3	9.3	3.5	3.3	3.6	5.4	8.0	5.1	4.2
State-space		CL	STS, local trend model, no correlation	7.6	14.5	7.0	8.5	6.5	6.2	6.4	7.3	8.9	8.3	8.9	9.1	8.3	8.0
			STS, smooth trend model, no correlation	6.8	14.6	5.4	7.1	6.0	5.5	7.1	7.6	9.4	8.7	6.6	11.2	8.0	7.9
			STS, local trend model, correlation	3.3	13.9	1.5	4.8	4.4	6.8	6.6	3.8	8.6	3.8	6.9	6.8	5.9	5.0
			STS, smooth trend model, correlation	6.1	14.2	5.2	6.9	4.8	5.6	7.3	6.4	10.6	6.5	7.3	8.8	7.5	7.1
		DC	STS, local trend model, no correlation	7.8	30.7	7.6	11.1	7.4	6.9	9.9	12.5	12.3	8.0	11.6	12.6	11.5	9.9
			STS, smooth trend model, no correlation	7.3	29.7	5.7	9.3	7.2	6.1	10.8	11.1	12.6	8.4	9.1	14.6	11.0	9.6
			STS, local trend model, correlation	4.0	30.2	1.8	8.7	4.6	7.3	11.1	6.8	11.3	4.6	8.8	10.4	9.1	7.0
			STS, smooth trend model, correlation	6.4	30.7	5.6	9.0	5.5	6.6	10.6	9.0	14.0	7.2	9.8	11.3	10.5	9.0

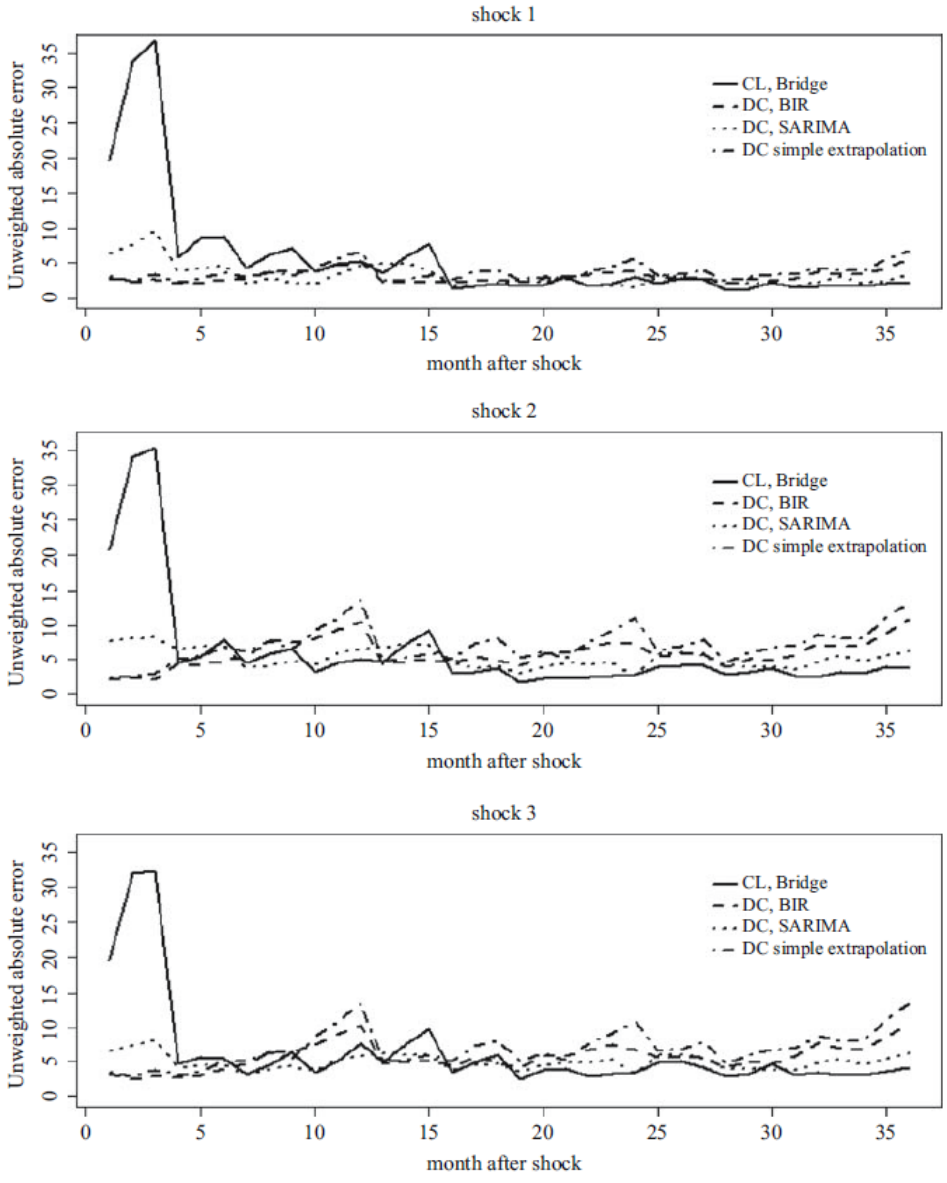
CL, Bridge model is again among the better performing nowcast models during this more stable period. We see that the NC models are less accurate than without the (simulated) crisis (also when we recalculate Table 5.1 over the same period January 2018 – December 2019, not shown). Of course, this is no surprise.

Table 5.7 shows that in year two and three after the shock, the CL, Bridge model is back among the best performing models, overtaking the models that did well in the first year after the crisis. The tables of shock 1 and 2 (not shown) give the same result. It is interesting to look more closely at the absolute error over time. Therefore, in Figure 5.4 below we show the average absolute error over all twelve PPCs, for month 1 to 36 after the shock occurred. Figure 5.4 shows that during the first three months the CL, Bridge model has much higher average absolute errors. This is a general pattern that holds for all NC before BM models (not shown in the figure). Another pattern in all three shocks is that during the first three months after a shock, the SE and BIR model perform best.

5.4 Conclusion

In this paper the estimation of short-term monthly estimates based on a slow but accurate quarterly series and a potentially selective monthly auxiliary series is discussed. There are two problems involved. First, the quarterly series must be temporally disaggregated (TD), using the monthly auxiliary series. This is done with well-known BM models. These models, Chow-Lin (CL) and Denton-Cholette (DC), transfer the monthly pattern of the monthly auxiliary series onto the quarterly series. Unfortunately, the plausibility of these transfers cannot be evaluated in this application, since the monthly patterns of the businesses that declare VAT on a quarterly frequency remains unknown. The fact, however, that subject matter specialists consider the results as plausible, gives us trust in the results. A major part of the paper concerns the second problem: monthly estimates must be computed before the quarterly figure is available, which means that a nowcast method must be applied. In the paper different nowcast methods are compared. In the evaluation of the methods, we distinguish between a stable economic period, where the development of the series is quite stable and predictable, and a period of crisis in which a sudden shock occurs. The financial crisis of 2008 and the COVID-19 pandemic are two examples of such crises, of which the latter is considered in this paper. The methods are applied to twelve series that are published by Statistics Netherlands. It is found that during a stable period most of the methods we consider perform quite well. The so-called Bridge, CL model performs slightly better than the other methods. This method first predicts both the quarterly and the monthly series (using a SARIMA-model) of the current quarter. Then, a CL BM model is applied. In a period of crisis, the Bridge model is no longer the most accurate model. However, during a shock most models perform worse than during a stable period. Right after a shock, NC after BM models perform better than the NC before BM models. The best method in a period of crisis seems to be the DC, SARIMA

Figure 5.4: Mean absolute error(MAE) over all PPCs, 1 – 36 months after occurrence of shock 1, 2 and 3, for selected models



model with the monthly auxiliary series as regressor and use this model to nowcast the current month. The reason why DC is preferred in times of crisis, is because an (S)ARIMA model of a series obtained by CL leads to larger dependencies on the history of the series. A simulation study shows that one year after a sudden shock, the CL, Bridge method is again one of the best performing methods. In this paper, for two reasons, only one auxiliary monthly series is used for both BM and nowcasting. The first reason is that each auxiliary monthly series is based on turnover of companies with similar economic activity (i.e. primary publication cells, PPCs). This implies that an auxiliary series that measures the same phenomenon is potentially sufficient, while additional auxiliary monthly series might introduce error. This may hold especially in a period of crisis, where the relation between the additional auxiliary series and the target series might be disturbed leading to model misspecification and biased nowcasts. The second reason is that in the production of timely monthly official figures there is very limited amount of time for model checking and evaluation. From that point of view, relatively simple models that are easy to interpret are preferred above complex models. In this paper twelve time series of so-called primary publication cells (PPCs) (e.g. 'Restaurants' or 'Publishers') are considered as test cases, and nowcasts are computed for 44 months in normal times and 4 months in crisis times. The performance of the different models is quite consistent over these test cases and within periods. This might indicate that our results can be generalized to other applications but more empirical results to support our findings are of course desired. An issue with the simulation in this study is to find a benchmark to evaluate the accuracy of the proposed methods. In this study the monthly index series as obtained by BM are used as the benchmark. It is not clear whether this choice favours some methods above others. A simulation that does not favour or handicap particular methods requires a setup where artificial populations are created. This indeed gives more insight in the properties of the different procedures under different conditions. This is left as further research. The index series which are used in this paper are based on turnover sums, where the monthly auxiliary series are based on a selective subpopulation. It is also possible to (partly) correct for this selectivity by weighting, using the available background information about the involved enterprises. In this application, only limited information is available. When there is some information about the self-selection process available, this could be used in the correction process as well. In this paper, we investigated the index series that are published by Statistics Netherlands as short-term statistics. Each index series is based on two turnover series. In a preliminary analysis it is investigated whether the accuracy of the nowcasts could be improved when these underlying turnover series are modelled instead of the index series. It is found that modelling the index series is more promising. See Zult, Krieg, Schouten, Ouwehand, and van den Brakel (2020) for more results and further details.

5.5 Appendix

A special structural time series model is developed to handle the different frequencies of the monthly and quarterly series. See Durbin and Koopman (2012, Ch. 6) for a general introduction and the estimation procedure of structural time series models. Whereas for the other methods the quarterly value is repeated three times in the quarterly series $y^Q_{Q(t)} = y^Q_{Q(t)+1} = y^Q_{Q(t)+2}$, for the STM the value of this series is missing in the first and second month of each quarter $y^Q_{Q(t)} = y^Q_{Q(t)+1} = \text{NA}$. STM can handle missings without problems. The time series $\mathbf{y_t} = \left(y^Q_t, x_t\right)^\top$ is modelled as:

$$\mathbf{y_t} = \mathbf{L_t} + \mathbf{S_t} + \mathbf{e_t}, \tag{5.9}$$

with $\mathbf{L_t} = (l^y_t, L^x_t)^\top$ the trend component, $\mathbf{S_t} = \left(s^y_t, S^x_t\right)^\top$ the seasonal component and $\mathbf{e}_t = \left(e^y_t, e^x_t\right)^\top$ the noise component. These three components are worked out as follows. For the trend, we consider 2 different models. The first STM model is shortly called the local trend model with correlation. In this case, the trend of the quarterly series is modelled as:

$$l^y_t = \left(L^y_{t-2} + L^y_{t-1} + L^y_t\right)/3, \tag{5.10}$$

And the underlying trend L^y_t and the trend of the monthly series L^x_t are modelled as

$$L^a_t = L^a_{t-1} + \eta^a_t \text{ with } a = \{x, y\},$$

$$\mathrm{E}[\eta^a_t] = 0, \tag{5.11}$$

$$\mathrm{Cov}(\eta^a_t, \eta^a_{t'}) = \begin{cases} \sigma^2_{L,a} & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases} \text{ and}$$

$$\mathrm{Cov}(\eta^y_t, \eta^x_{t'}) = \begin{cases} \zeta^2_L & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases}$$

The second STM model is called the smooth trend model with correlation. The trend of the quarterly series is again modelled with Eq. (5.10), but now, L^y_t and L^x_t are modelled as

$$L^a_t = L^a_{t-1} + R^a_{t-1},$$

$$R^a_t = R^a_{t-1} + \eta^a_t, \text{ with } a = \{x, y\}, \tag{5.12}$$

The variance – covariance structure of η_t^a is as in Eq. (5.11). Remark: The trend of the quarterly series is the mean of the trend of the monthly series in three consecutive months, modelled with Eq. (5.10). This is only relevant for the third month of every quarter, as for the other months the model computes a trend L_t^y and l_t^y for the first and second month of each quarter. The seasonal component of the monthly series is modelled with the well-known trigonometric seasonal model (for monthly figures). See Durbin and Koopman (2012) for details. The details of the seasonal component of the quarterly series are explained in the main paper. The noise component is modelled with white noise. To consider that both series are based partly on the same enterprises, two independent white noise variables are modelled:

$$\mathrm{E}\left[\epsilon_{t,j}\right] = 0,$$

$$\mathrm{Cov}(\epsilon_{t,j}, \epsilon_{t',j}) = \begin{cases} \sigma_{\epsilon,j}^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases} \text{ with } j = 1, 2 \tag{5.13}$$

$$\mathrm{Cov}(\epsilon_{t,1}, \epsilon_{t',2}) = 0.$$

Then, $e_t^x = \epsilon_{t,1} + \epsilon_{t,2}$, and

$$e_t^y = \begin{cases} \frac{\epsilon_{t-2,1}+\epsilon_{t-1,1}+\epsilon_{t,1}}{3} & \text{if } t \text{ is the third month of the quarter} \\ 0 & \text{if } t \text{ is the first or second month of the quarter} \end{cases}$$

In the current presentation, the sum of two independent white noise processes is used to model the measurement error in the monthly figures. The quarterly average of one of them is shared by the quarterly series. In this way the model accounts for the correlation in the measurement disturbance terms of the monthly and quarterly series, since the monthly and quarterly declarants are both used when the quarterly series is computed. Alternatively, it is also possible to add the two independent white noise components to the quarterly series where one of them is shared by the monthly series. Since the process of the time series components are defined with a monthly frequency it is more straightforward to add the two white noise components to the monthly series. The analysis is conducted with software developed in OxMetrics in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman, Shephard, and Doornik (2008).

Chapter 6

Nowcasting in triple-system estimation

When samples that each cover part of a population for a certain reference date become available slowly over time, an estimate of the population size can be obtained when at least two samples are available. Ideally one uses all the available samples, but if some samples become available much later one may want to use the samples that are available earlier, to obtain a preliminary or nowcast estimate. However, a limited number of samples may no longer lead to asymptotically unbiased estimates, in particular in the case of two early available samples that suffer from pairwise dependence. In this paper we propose a multiple system nowcasting model that deals with this issue by combining the early available samples with samples from a previous reference date and the expectation-maximisation algorithm. This leads to a nowcast estimate that is asymptotically unbiased under more relaxed assumptions than the dual-system estimator. The multiple system nowcasting model is applied to the problem of estimating the number of homeless people in The Netherlands, which leads to reasonably accurate nowcast estimates.

This chapter is in a revision process (status: accepted under minor revision) for publication. A preliminary version is available at arXiv, https://arxiv.org/abs/2406.17637, (Zult, van der Heijden, & Bakker, 2024). Author contributions: DZ and PvdH posed the problem and worked out the idea. DZ did the analyses and wrote the text. BB and PvdH discussed and edited the text.

6.1 Introduction

A well-known problem in the production of statistics is that data may become available gradually, while a statistic for a certain reference date has to be produced before all these data are available. In such cases, it is common practice to produce a preliminary statistic that can also be referred to as a nowcast, based on the data that is available at the time of publication, and update this statistic shortly after the delivery date of the last sample. Discussions on this topic usually evolve around correcting for response bias that may occur when the speed of response is related to the statistic itself. For example, when companies with a quickly growing turnover also respond quickly, a nowcast on turnover growth might be biased upwards if this relation is ignored.

A statistic for which such a nowcasting method is not available, is a population size estimate based on samples that each partly observe a population, and where one or more complete samples are available with delay. This may occur when, for example, samples are registers or surveys that are maintained or collected periodically throughout a certain period. Then, some samples might be available early and others later, although they refer to the same reference date. In such cases it is common practice to simply wait until all samples have become available before estimation is performed. This raises the question whether and under what conditions it is possible to produce a preliminary population size estimate based on the set of samples that are available earlier. The most simple case is when for the reference date one sample becomes available earlier and a second sample becomes available later. A slightly more complex case is when for a reference date three samples become available sequentially with some time in between, which is the main topic of this paper.

The models that are involved in the estimation of the size of a partly observed population are known under different names such as capture-recapture, mark and recapture or multiple systems estimation (MSE). When the number of samples is two or three, MSE is usually referred to as dual-system estimation (DSE) or triple-system estimation (TSE), respectively. The most basic DSE model was proposed by Petersen (1896), and later by Lincoln (1930). Under a set of assumptions discussed by Wolter (1986), their DSE estimator provides an asymptotically unbiased population size estimate. A DSE assumption that is often unlikely to hold, is the independence of the two samples. This independence assumption can be relaxed when three or more samples are available, and therefore, as discussed by Fienberg (1972), TSE is often recommended.

The case considered in this paper is that a contingency table based on three samples for the previous reference date, and a contingency table based on one or two samples for the current reference date is available. The goal is to obtain a maximum likelihood (ML) population size estimate for the current reference date. The absence of a second and third or only a third sample for the current reference date could be considered a missing data problem. A standard method to deal with this issue is the expectation–maximization (EM) algorithm (see e.g. Dempster et al., 1977). The EM

algorithm method allows for statistical inference from incomplete data with ML. In this paper we will discuss under which conditions the EM algorithm can be combined with DSE and TSE to obtain an asymptotically unbiased preliminary population size estimate, which we will refer to as nowcast (NC) estimate. This approach of combining the EM algorithm with MSE models based on incomplete data is not new. For example, Zwane, van der Pal-de Bruin, and van der Heijden (2004) consider the case that some samples may contain different but overlapping populations, and Zwane and van der Heijden (2007) consider the case where some covariates are missing in some samples. New in this study is that the method is applied to obtain nowcasts for which both observations and estimates based on fully observed MSE data become available later. This allows us to compare the nowcasting model estimates with actual observations and the estimate based on fully observed MSE data in a practical example.

Next, Section 6.2 discusses the DSE and TSE model, and how data for two periods can be combined in one framework. This framework contains incomplete data, therefore Section 6.3 discusses how the EM algorithm can be used to obtain ML estimates from this framework. This combination of DSE, TSE and the EM algorithm gives a MSE nowcasting model. Finally, in Section 6.4 we will apply this model to obtain nowcasts for the number of homeless people in The Netherlands, and compare these nowcasts with alternative estimates such as the standard DSE estimate.

6.2 Theory and notation

This section discusses DSE and TSE notation and theory, and shows how DSE and TSE models can be combined over two periods.

6.2.1 Dual-system estimation

Imagine a population with size N and a set of two samples A and B that each cover part of this population. The goal is to use these samples to obtain a population size estimate denoted as $\hat{N}$. When each unit in each sample can be uniquely identified, then for each unit an inclusion pattern ab can be constructed, with $a, b \in \{1, 0\}$, where $a = 1$ stands for "*included in sample A*" and $a = 0$ for "*not included in sample A*", and the same with b for sample B. The units of each inclusion pattern can be counted and denoted as n_{ab}, except when the inclusion pattern is 00, because these units are unobserved. The sum of all observed units is denoted as n and so $n = n_{11} + n_{10} + n_{01}$. Finally, when we sum over a or b, we replace that subscript by a "+". Thus, for example, $n_{1+} = n_{10} + n_{11}$ is equal to the size of source A. It is assumed that n_{ab} is a realisation of a random variable with expectation m_{ab} and the aim of DSE is to obtain $\hat{m}_{ab}$, an estimate of this expectation.

Under a set of assumptions discussed by for example Wolter (1986), the observed counts n_{11}, n_{10} and n_{01} can be used to estimate N. These assumptions can be summarised as:

1. The true population is equal for the samples A and B.
2. Records that correspond to the same unit in sample A and B can be perfectly linked.
3. Inclusion probabilities are homogeneous in sample A or B (see e.g. Seber, 1982).
4. Sample A and B are independent.

Under assumption (1-4), an asymptotically unbiased DSE-estimator for m_{00} can be written as

$$\hat{m}_{00}^{\text{DSE}} = \frac{n_{10} n_{01}}{n_{11}}, \tag{6.1}$$

and consequently for N as $\hat{N}^{\text{DSE}} = n + \hat{m}_{00}^{\text{DSE}} = \frac{n_{1+} n_{+1}}{n_{11}}$.

Fienberg (1972) showed that the DSE estimator can also be derived from a log-linear model for m_{ab}, and for our purpose it is important to show how this relates to the independence assumption 4. A log-linear model for m_{ab} can be written as

$$\log m_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB}, \tag{6.2}$$

with λ an intercept term, λ_a^A and λ_b^B are the respective inclusion parameters for sample A and B that are identified by setting $\lambda_0^A = \lambda_0^B = 0$ and λ_{ab}^{AB} is a parameter for the interaction between sample A and B. Because m_{00} is unobserved and the independence assumption 4 implies that $\lambda_{ab}^{AB} = 0$, in practice Eq. (6.2) represents three equations and three unknowns that lead to the DSE-estimator in Eq. (6.1). This also shows that if $\lambda_{ab}^{AB} \neq 0$, then $\hat{m}_{00}^{\text{DSE}}$ is a biased estimate for m_{00}. In the next section we will show how TSE may solve this problem of bias due to pairwise dependence of samples.

6.2.2 Triple-system estimation

When instead of two samples, a population is partly observed by three samples A, B and C, each unit has an inclusion pattern that, instead of ab, can be written as abc, where c is defined in the same way as a and b. This means that instead of the four inclusion patterns in DSE there are now eight TSE inclusion patterns 000, 100, 010, 001, 110, 101, 011 and 111, and Eq. (6.2) can be extended towards

$$\log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ab}^{AB} + \mu_{ac}^{AC} + \mu_{bc}^{BC} + \mu_{abc}^{ABC}. \tag{6.3}$$

Eq. (6.3) constitutes a system of eight linear equations and eight unknowns, but because m_{000} is unknown, it cannot be solved. Therefore it is usually assumed that $\mu_{abc}^{ABC} = 0$, which is similar but more realistic than DSE assumption 4. This assumption gives the so-called saturated TSE model

$$\text{saturated: } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ab}^{AB} + \mu_{ac}^{AC} + \mu_{bc}^{BC}, \tag{6.4}$$

that in contrast to DSE, also contains pairwise interaction parameters μ_{ab}^{AB}, μ_{ac}^{AC} and μ_{bc}^{BC}. This model can be further restricted by setting one or more pairwise interaction terms to zero, which gives seven additional models, i.e.:

$$\text{two-pair dependence (I): } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ac}^{AC} + \mu_{bc}^{BC}, \tag{6.5}$$

$$\text{two-pair dependence (II): } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ab}^{AB} + \mu_{bc}^{BC}, \tag{6.6}$$

$$\text{two-pair dependence (III): } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ab}^{AB} + \mu_{bc}^{AC}, \tag{6.7}$$

$$\text{one-pair dependence (I): } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{bc}^{BC}, \tag{6.8}$$

$$\text{one-pair dependence (II): } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ac}^{AC}, \tag{6.9}$$

$$\text{one-pair dependence (III): } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C + \mu_{ab}^{AB}, \tag{6.10}$$

$$\text{independence: } \log m_{abc} = \mu + \mu_a^A + \mu_b^B + \mu_c^C. \tag{6.11}$$

Making the distinction between these restricted models is important when TSE and DSE over two periods is combined. This will be discussed in the next section. Models with more than three samples can be developed along the same lines.

6.2.3 Combining samples over two periods.

We consider a population with size N_t and the samples A_t, B_t and C_t that each cover parts of this population for reference date t. Also assume the delivery dates $t = t_0, t_{1,a}, t_{1,b}, t_{1,c}$ where at t_0 the samples A_{t_0}, B_{t_0} and C_{t_0} for reference date $t = t_0$ are all available and at delivery dates $t_{1,a}, t_{1,b}$ and $t_{1,c}$ the samples A_{t_1}, B_{t_1} and C_{t_1} for reference date $t = t_1$ become available, one-by-one, in that order. This means that at both $t = t_0$ and $t = t_{1,c}$ three samples are available for their corresponding periods t_0 and t_1. When we write abc, t as the inclusion pattern for reference date t, a table can be constructed that shows which observed counts are available at which moment, as in Table 6.1 below.

Table 6.1 shows that for $t = t_0$ and $t = t_{1,c}$ all observed counts are available for their corresponding reference dates, and so for each reference date a TSE-estimate for $m_{000,t}$, as discussed in Section 6.2.2, can be estimated. We write their corresponding TSE models as $M_{t_0}(\boldsymbol{\mu}_{t_0})$ and $M_{t_{1,c}}(\boldsymbol{\mu}_{t_{1,c}}) = M_{t_1}(\boldsymbol{\mu}_{t_1})$ with $\boldsymbol{\mu}_t$ as the vector of μ_t-parameters at reference date t. At $t = t_{1,a}$ and $t = t_{1,b}$ this is not possible, because at those delivery dates only one or two samples are available for reference date t_1. Table 6.1 shows that at those moments only aggregated observed counts are available. Then the question becomes if and under which assumptions, the old samples A_{t_0}, B_{t_0} and C_{t_0}, together with these aggregated observed counts, can be used to obtain an asymptotically unbiased estimate for N_{t_1}. In general, for each observed count that corresponds to a reference date t, one additional parameter for that reference date can be estimated. This reasoning allows us to construct MSE models for the case that samples correspond to different reference dates.

At $t = t_{1,a}$ the additional observed count n_{1++,t_1} becomes available, which simply is the total sample size of A_{t_1}. This can be considered one observed count for reference

Table 6.1: Combined table at $t = t_0$, $t_{1,a}$, $t_{1,b}$ and $t_{1,c}$.

A	B	C	t	n_{abc,t_0}	$n_{abc,t_{1,a}}$	$n_{abc,t_{1,b}}$	$n_{abc,t_{1,c}}$
1	1	1	t_0	n_{111,t_0}	n_{111,t_0}	n_{111,t_0}	n_{111,t_0}
1	1	0	t_0	n_{110,t_0}	n_{110,t_0}	n_{110,t_0}	n_{110,t_0}
1	0	1	t_0	n_{101,t_0}	n_{101,t_0}	n_{101,t_0}	n_{101,t_0}
1	0	0	t_0	n_{100,t_0}	n_{100,t_0}	n_{100,t_0}	n_{100,t_0}
0	1	1	t_0	n_{011,t_0}	n_{011,t_0}	n_{011,t_0}	n_{011,t_0}
0	1	0	t_0	n_{010,t_0}	n_{010,t_0}	n_{010,t_0}	n_{010,t_0}
0	0	1	t_0	n_{001,t_0}	n_{001,t_0}	n_{001,t_0}	n_{001,t_0}
0	0	0	t_0	?	?	?	?
1	1	1	t_1	?		n_{11+,t_1}	n_{111,t_1}
1	1	0	t_1	?			n_{110,t_1}
1	0	1	t_1	?	n_{1++,t_1}		n_{101,t_1}
1	0	0	t_1	?		n_{10+,t_1}	n_{100,t_1}
0	1	1	t_1	?	?		n_{011,t_1}
0	1	0	t_1	?	?	n_{01+,t_1}	n_{010,t_1}
0	0	1	t_1	?	?	?	n_{001,t_1}
0	0	0	t_1	?	?	?	?

date $t = t_1$ and therefore allows a model with one additional parameter for reference date $t = t_1$, i.e.

$$M_{t_{1,a}}(\boldsymbol{\mu_{t_{1,a}}}) = \log m_{abc,t} = M_{t_0}(\boldsymbol{\mu_{t_0}}) + \mu_{t_1}, \tag{6.12}$$

where $M_{t_0}(\boldsymbol{\mu_{t_0}})$ is one of the models in Eq. (6.4 - 6.11) with t_0 attached in each subscript of each μ-parameter. Note that the parameter μ_{t_1} is an additional constant that is added to μ_{t_0} in case of reference date t_1, so for m_{000,t_1}, Eq. (6.12) reduces to the expression $m_{000,t_1} = \exp\left(\mu_{t_0} + \mu_{t_1}\right)$. The remaining parameters in $M_{t_0}(\boldsymbol{\mu_{t_0}})$ are assumed to hold for both reference dates t_0 and t_1. The ML estimate for μ_{t_0} is assumed to be asymptotically unbiased if model $M_{t_0}(\boldsymbol{\mu_{t_0}})$ is true, but whether the ML estimate for μ_{t_1} is also asymptotically unbiased depends on the remaining parameters in $M_{t_0}(\boldsymbol{\mu_{t_0}})$. If inclusion probabilities in and pairwise dependencies between sample A_t, B_t and C_t are independent of t, the ML-estimators for the remaining parameters are asymptotically unbiased estimators for both reference dates, and then the ML-estimator for μ_{t_1} is also an asymptotically unbiased estimator. In that case the ML-estimator for m_{000,t_1} and therefore N_{t_1} is an asymptotically unbiased estimator too.

At $t = t_{1,b}$ the additional sample B_{t_1} becomes available and so at $t = t_{1,b}$ two samples are available for reference date $t = t_1$. Table 6.1 shows that this means that three observed counts, with inclusion patterns $abc = 11+, 10+, 01+$, are available for this reference date. This implies that for reference date $t = t_1$ a DSE-estimate can be obtained, but as was discussed in Section 6.2.1, this estimate is biased if the independence assumption is violated. Then the question becomes if the presence of the samples A_{t_0},

B_{t_0} and C_{t_0} allows for a way in which the independence assumption can be relaxed. Note that due to the three observed counts we can extend $M_{t_{1,a}}(\boldsymbol{\mu_{t_{1,a}}})$ in Eq. (6.12) with two additional parameters for $t = t_1$, i.e.

$$M_{t_{1,b}}(\boldsymbol{\mu_{t_{1,b}}}) = \log m_{abc,t} = M_{t_0}(\boldsymbol{\mu_{t_0}}) + \mu_{t_1} + \mu^A_{a,t_1} + \mu^B_{b,t_1}. \tag{6.13}$$

This model gives the same expression $\exp\left(\mu_{t_0} + \mu_{t_1}\right)$ for m_{000,t_1} as $M(t_{1,a})$, but the conditions under which the ML-estimator for the parameter μ_{t_1} is an asymptotically unbiased estimator are more relaxed. Note that the remaining parameters in $M_{t_0}(\boldsymbol{\mu_{t_0}})$ that are assumed to hold for both periods have reduced with μ^A_{a,t_0} and μ^B_{b,t_0}, which now, due to the presence of μ^A_{a,t_1} and μ^B_{b,t_1}, correspond exclusively to inclusion probabilities for reference date t_0. Therefore, for model $M_{t_{1,b}}(\boldsymbol{\mu_{t_{1,b}}})$ to hold, as compared to model $M_{t_{1,a}}(\boldsymbol{\mu_{t_{1,a}}})$, a reduced set of remaining parameters in $M_{t_0}(\boldsymbol{\mu_{t_0}})$ should be independent of t. This implies that in model $M_{t_{1,b}}(\boldsymbol{\mu_{t_{1,b}}})$ the inclusion probabilities for sample A_{t_1} and B_{t_1} may differ from the inclusion probabilities for sample A_{t_0} and B_{t_0}.

Finally, it is instructive to compare Eq. (6.13) with the DSE Eq. (6.2). When $m_{abc,t} = m_{ab}$, $\mu_{t_1} = \lambda$, $\mu^A_{a,t_1} = \lambda^A_a$, $\mu^B_{b,t_1} = \lambda^B_b$ and $M_{t_0}(\boldsymbol{\mu_{t_0}}) = \lambda^{AB}_{ab}$, the equations are equivalent. This implies that for $M_{t_{1,b}}(\boldsymbol{\mu_{t_{1,b}}})$ the DSE independence assumption 4. can be replaced by the (more relaxed) assumption

4. The pairwise dependence parameter λ^{AB}_{ab} is independent of t.

In other words, the estimate for λ^{AB}_{ab} for the previous reference date can be used as an estimate for the current reference date, because it is assumed to be stable between both periods.

The estimation of the parameters in the models $M_{t_{1,a}}(\boldsymbol{\mu_{t_{1,a}}})$ and $M_{t_{1,b}}(\boldsymbol{\mu_{t_{1,b}}})$ is less straightforward than the estimation of the parameters in $M_{t_0}(\boldsymbol{\mu_{t_0}})$ and $M_{t_1}(\boldsymbol{\mu_{t_1}})$, which can be estimated directly with ML. How to deal with this problem is discussed in the next section.

6.3 Combining DSE and TSE with the EM algorithm

Table 6.1 from the previous section poses two statistical estimation problems. On top of the problem of the unobserved counts n_{000,t_0} and n_{000,t_1}, it also poses a so-called mixture model problem (see e.g. Lindsay, 1995). This problem implies that for (some) variables only an aggregate over different groups is observed, or one may say that for some groups the data is incomplete. In this case, at $t = t_{1,a}$, there is the aggregated observed count $n_{1{+}{+},t_1}$ and at $t = t_{1,b}$ there are the three aggregated observed counts $\left(n_{11{+},t_1}, n_{10{+},t_1}, n_{01{+},t_1}\right)$. $n_{1{+}{+},t_1}$ is simply the size of sample A_{t_1}, and $\left(n_{11{+},t_1}, n_{10{+},t_1}, n_{01{+},t_1}\right)$ are the aggregated observed counts over sample C_{t_1} of the units included in sample A_{t_1} and/or B_{t_1}. A standard method to deal with incomplete data is the EM algorithm. In this case it allows for the estimation of the underlying counts

that together add up to the observed aggregated counts, such as the unobserved n_{111,t_1} and n_{110,t_1} at $t = t_{1,b}$ that add up to the observed n_{11+,t_1}.

The EM algorithm was introduced by Dempster et al. (1977) as a tool to obtain ML-estimates in case of incomplete data due to unobserved or latent variables. In the problem discussed in this paper, the EM algorithm can be applied with model $M_{t_{1,a}}(\boldsymbol{\mu}_{\boldsymbol{t_{1,a}}})$ or $M_{t_{1,b}}(\boldsymbol{\mu}_{\boldsymbol{t_{1,b}}})$ in Eq. (6.12) and (6.13). For this case, the outcome of the EM algorithm at $t = t_{t,a}$ and $t = t_{t,b}$ is shown in Table 6.2.

Table 6.2: Table with completed data.

A	B	C	t	$\hat{n}_{abc,t_{1,a}}$	$\hat{n}_{abc,t_{1,b}}$
1	1	1	t_0	n_{111,t_0}	n_{111,t_0}
1	1	0	t_0	n_{110,t_0}	n_{110,t_0}
1	0	1	t_0	n_{101,t_0}	n_{101,t_0}
1	0	0	t_0	n_{100,t_0}	n_{100,t_0}
0	1	1	t_0	n_{011,t_0}	n_{011,t_0}
0	1	0	t_0	n_{010,t_0}	n_{010,t_0}
0	0	1	t_0	n_{001,t_0}	n_{001,t_0}
0	0	0	t_0	?	?
1	1	1	t_1	$\hat{n}_{111,t_{1,a}}$	$\hat{n}_{111,t_{1,b}}$
1	1	0	t_1	$\hat{n}_{110,t_{1,a}}$	$\hat{n}_{110,t_{1,b}}$
1	0	1	t_1	$\hat{n}_{101,t_{1,a}}$	$\hat{n}_{101,t_{1,b}}$
1	0	0	t_1	$\hat{n}_{100,t_{1,a}}$	$\hat{n}_{100,t_{1,b}}$
0	1	1	t_1	?	$\hat{n}_{011,t_{1,b}}$
0	1	0	t_1	?	$\hat{n}_{010,t_{1,b}}$
0	0	1	t_1	?	?
0	0	0	t_1	?	?

To illustrate how the Expectation step (E-step) of the EM algorithm yields completed data in the columns $\hat{n}_{abc,t_{1,a}}$ and $\hat{n}_{abc,t_{1,b}}$ in Table 6.2, we discuss this for $\hat{n}_{abc,t_{1,b}}$. The EM algorithm allows to split-up n_{ab+,t_1} into the completed data $\hat{n}_{ab1,t_{1,b}}$ and $\hat{n}_{ab0,t_{1,b}}$ with $\hat{n}_{ab1,t_{1,b}} + \hat{n}_{ab0,t_{1,b}} = n_{ab+,t_1}$. The EM algorithm starts with an initialisation step that creates an initial set of completed data by, for example, $\hat{n}^{(0)}_{ab1,t_{1,b}} = n_{ab+,t_1}/2$ and $\hat{n}^{(0)}_{ab0,t_{1,b}} = n_{ab+,t_1}/2$. Next, in the first maximisation step (M-step) these completed data are assumed regular observations that, together with n_{abc,t_0}, can be used to estimate the parameters of the model $M_{t_{1,b}}(\boldsymbol{\mu}_{\boldsymbol{t_{1,b}}})$ in Eq. (6.13), but here it is also possible to replace $M_{t_0}(\boldsymbol{\mu}_{\boldsymbol{t_0}})$ with a more restricted model. The model resulting from this M-step gives, at iteration 0, the fitted values $\hat{m}^{(0)}_{abc,t}$. Next, in the first expectation step (E-step) these fitted values are used to (again) split-up n_{ab+,t_1}, but now as

$\hat{n}^{(1)}_{ab1,t_{1,b}} = n_{ab+,t_1}(\hat{m}^{(0)}_{ab1,t_{1,b}}/\hat{m}^{(0)}_{ab+,t_{1,b}})$ and $\hat{n}^{(1)}_{ab0,t_{1,b}} = n_{ab+,t_1}(\hat{m}^{(0)}_{ab0,t_{1,b}}/\hat{m}^{(0)}_{ab+,t_{1,b}})$, which gives a new set of completed data that can be used to, again, estimate the model $M_{t_{1,b}}(\boldsymbol{\mu}_{\boldsymbol{t_{1,b}}})$ in Eq. (6.13). This iterative procedure repeats itself i times until $\hat{n}^{(i)}_{abc,t_{1,b}}$ converges. The resulting set of stabilised completed data are the $\hat{n}_{abc,t_{1,b}}$ in Table 6.2, and they are used to derive maximum likelihood estimates $\hat{m}_{abc,t_{1,b}}$.

The last M-step provides fitted values $\hat{m}_{abc,t}$ for each cell, including the cells with inclusion patterns $001, t_1$ and $000, t_1$. We refer to these estimates as $\hat{m}^{\text{NC}}_{abc,t}$ and summing up over them for $t = t_{1,b}$ gives a fitted value for N_{t_1}. We refer to this sum as the nowcast estimate for N_{t_1}, i.e.

$$\hat{N}^{\text{NC}}_{t_1} = \sum_{abc \in ABC} \hat{m}^{\text{NC}}_{abc,t_{1,b}}, \tag{6.14}$$

with ABC the set of all inclusion patterns. In the next section we will use this estimator to obtain nowcasts for the number of homeless people in The Netherlands.

6.4 Nowcasting the number of homeless people in The Netherlands

In this section we investigate how the MSE nowcasting model performs by using a dataset that is also used to estimate the number of homeless people in The Netherlands. The estimation of the number of homeless people in The Netherlands is discussed in detail in Coumans et al. (2017). The estimation procedure is based on three samples that we refer to as sample A_y, B_y and C_y, where y indicates the year, and is performed annually. The resulting TSE estimate for the 1^{st} of January of each year is based on a model selection procedure that leads to a TSE model that also includes a set of covariates, namely sex, age, region of stay and region of birth. The samples that are used become available over a year, where the first two samples A_y and B_y are available early during the year and the third sample C_y is available somewhere in the third or fourth quarter of the year. Data is available for each year over the period 2010 – 2023, except for the COVID-19 year 2019. The sample size for each sample in each year is presented in Table 6.3 below.

The scheme in which the samples become available implies that at $y = y_{t_{1,b}}$, for the years 2011 – 2018 and 2021 – 2023, both a DSE estimate and a NC estimate can be obtained. The fact that a NC estimate, as discussed in Section 6.2.3 and defined in Eq. (6.14), requires samples from two consecutive years means that it cannot be calculated for the years 2010, 2019 and 2020, because in those years data for the previous or next year are missing.

To simplify the interpretability of the results, both the model selection procedure is skipped by assuming a saturated model and the covariates are ignored by aggregating over them. Ignoring the covariates simplifies the data in Table 6.1 in Section 6.2.3. Second, skipping the model selection procedure and simply assuming the saturated

Table 6.3: Sample size for each year

Year	Sample size A_y	Sample size B_y	Sample size C_y
2010	2916	1746	3494
2011	3058	1644	3812
2012	2594	1505	3459
2013	2703	1491	3876
2014	2380	1566	4267
2015	2232	1475	4669
2016	2631	1130	5220
2017	2502	1139	5611
2018	2456	927	5824
2019	NA	NA	NA
2020	1928	2501	5808
2021	1992	2827	6213
2022	2371	2263	5018
2023	2554	3017	4315

model in Eq. (6.4) for each reference date, allows for a more straightforward comparison of the resulting estimates, because they cannot differ due to different models selected for different reference dates.

To further increase the generality of the analysis the order in which the samples become available is varied. In reality sample C_y is available last, but for analytical purposes this could as well be assumed to be sample A_y or B_y. The samples for reference date of year y that are used in the calculation of an estimate are given as additional information in the subscript. For example, a NC estimate based on sample A_{y-1}, B_{y-1}, C_{y-1}, A_y and B_y but not C_y, is denoted as $\hat{N}^{\text{NC}}_{ab,y}$.

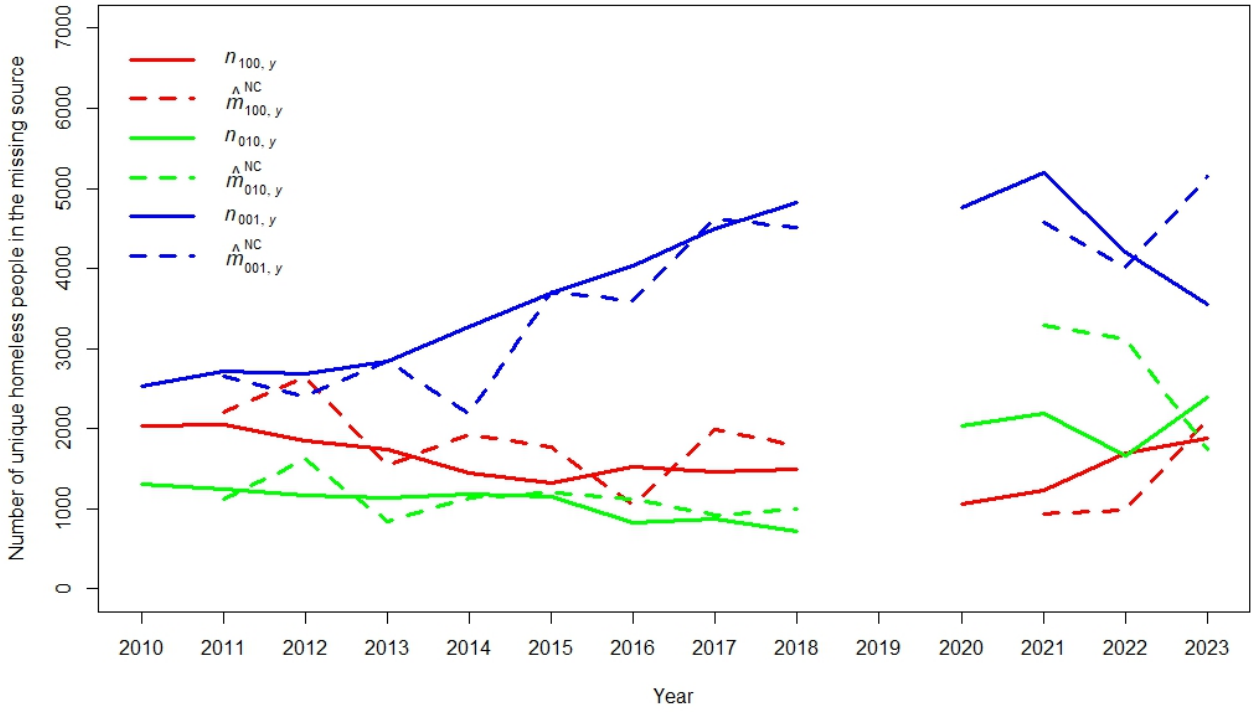
6.4.1 Results

This section presents the nowcasting model results for the homeless data. The results of the nowcasting model are evaluated in three ways. First, the nowcasting estimates for $m_{001,y}$ are compared with the actually observed $n_{001,y}$. Second, the time series of estimates for $\mu^{AB}_{ab,y}$, $\mu^{AC}_{ac,y}$ and $\mu^{BC}_{bc,y}$ are presented, which shows whether the nowcasting model assumption of stability of pairwise-dependencies between two periods is reasonable. Finally, the nowcasting model estimates for N_y are compared with the TSE model estimates for N_y.

Figure 6.1 shows the observed ($n_{100,y}$, $n_{010,y}$ and $n_{001,y}$) and nowcasting model estimates for the expected number of homeless people ($\hat{m}^{\text{NC}}_{100,y}$, $\hat{m}^{\text{NC}}_{010,y}$ and $\hat{m}^{\text{NC}}_{001,y}$) in the sample that is unavailable. Here the recent sample that is unavailable in the nowcasting model is indicated by the position of the "1" in the inclusion pattern in the

subscript. For example, $\hat{m}^{\text{NC}}_{001,y}$ is a nowcast that is based on sample A_y and B_y and not C_y. These nowcasting model estimates are interesting because they can be directly compared with observed values, which is rare in MSE models, because true population sizes generally remain unknown. The solid lines represent a series of observed counts and the dotted lines with corresponding colors represent the corresponding nowcasting model estimates. Figure 6.1 shows that irrespective of the unavailable

Figure 6.1: Observations and nowcasts of the number of homeless people that are uniquely observed in the missing sample over the periods 2010-2018 and 2020-2023.



sample, the nowcasting model estimates $\hat{m}^{\text{NC}}_{100,y}$, $\hat{m}^{\text{NC}}_{010,y}$ and $\hat{m}^{\text{NC}}_{001,y}$ follow a similar trend as the observed counts n_{100}, n_{010} and n_{001} that are available later, although for some year/missing sample combinations the difference can be quite substantial, especially the difference between the green solid and green dotted line for the years 2021 and 2022 stands out.

A similar figure can be constructed with a time series of TSE estimates ($\hat{N}_y^{\text{TSE}}$) based on all samples and the DSE ($\hat{N}_{bc,y}^{\text{DSE}}$, $\hat{N}_{ac,y}^{\text{DSE}}$ and $\hat{N}_{ab,y}^{\text{DSE}}$) and NC ($\hat{N}_{bc,y}^{\text{NC}}$, $\hat{N}_{ac,y}^{\text{NC}}$ and $\hat{N}_{ab,y}^{\text{NC}}$) estimates based on early available samples. The samples that are used in the estimation are indicated in the subscripts. For example, $\hat{N}_{ab,}^{\text{DSE}}$ and $\hat{N}_{ab,}^{\text{NC}}$ are a DSE and NC estimate based on sample A_{t_1} and B_{t_1}, while C_{t_1} is missing. These series are presented in Figure 6.2 below.

Figure 6.2: Estimates of the number of the total number of homeless people in The Netherlands over the periods 2010 – 2018 and 2020 – 2023.

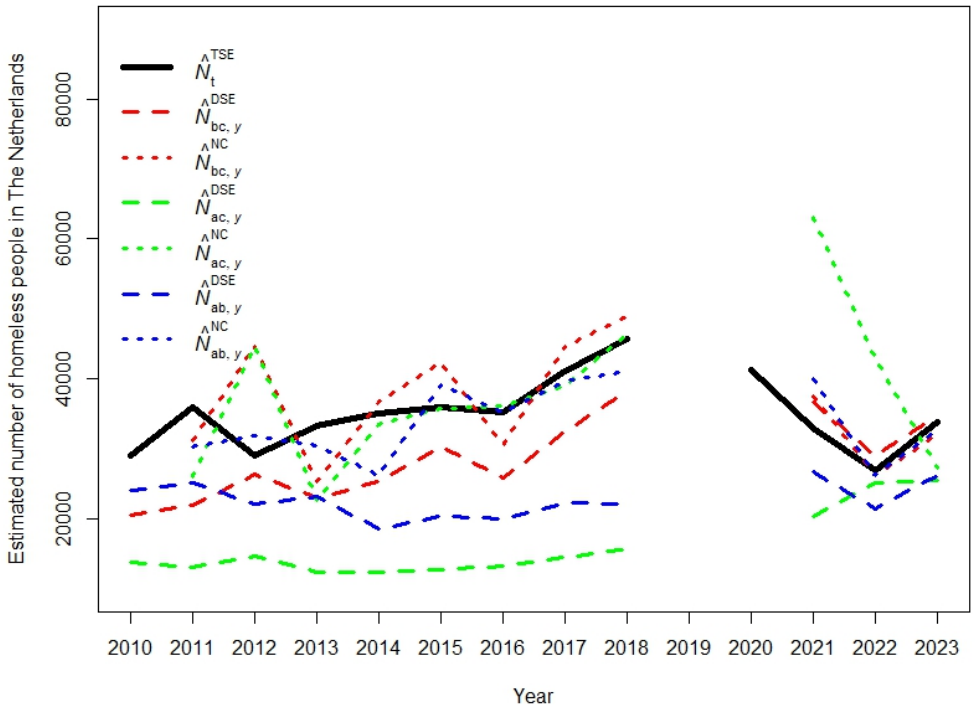


Figure 6.2 shows that for most years the nowcasting model estimates are much closer to the TSE estimates than the DSE estimates, which suggest that in this case the nowcasting model assumption of $\lambda_{ab,(y-1)}^{AB} = \lambda_{ab,y}^{AB}$ is more reasonable than the DSE assumption $\lambda_{ab,(y-1)}^{AB} = 0$. However, for some years the nowcasting model estimate can be quite bad, such as $\hat{N}_{ac,y}^{\text{NC}}$ in the years 2021 and 2022.

For many years it is questionable if the nowcasting model estimate is a better estimate than the TSE estimate of the previous year. In such cases a nowcast has no clear value added. To look deeper into this issue, Table 6.4 presents the differences between the TSE estimates and the lagged TSE estimates and nowcasting model estimates. Table 6.4 shows that the proximity of the nowcasting model estimates and the TSE estimate clearly differs for each sample delivery order. The best results are in the last column $\hat{N}_{ab,y}^{\text{NC}} - \hat{N}_y^{\text{TSE}}$, which has the lowest mean absolute difference (3.3), which

Table 6.4: Difference per year (×1000) between the TSE estimate and different estimates for each year

Year	$\hat{N}^{\text{TSE}}_{(y-1)} - \hat{N}^{\text{TSE}}_{y}$	$\hat{N}^{\text{NC}}_{bc,y} - \hat{N}^{\text{TSE}}_{y}$	$\hat{N}^{\text{NC}}_{ac,y} - \hat{N}^{\text{TSE}}_{y}$	$\hat{N}^{\text{NC}}_{ab,y} - \hat{N}^{\text{TSE}}_{y}$
2011	-6.9	-4.8	-9.9	-5.7
2012	-7.0	15.7	15.6	2.8
2013	4.3	-8.1	-10.7	-2.8
2014	1.8	1.6	-1.6	-8.9
2015	0.9	6.3	-0.2	3.1
2016	-0.8	-4.7	0.9	0.0
2017	6.0	3.5	-2.2	-1.4
2018	4.6	3.2	0.6	-4.7
2021	-8.3	4.6	30.2	7.2
2022	-6.0	-0.7	16.1	-0.6
2023	6.9	-1.8	-6.5	-0.9
Mean absolute difference	4.5	4.7	8.1	3.3

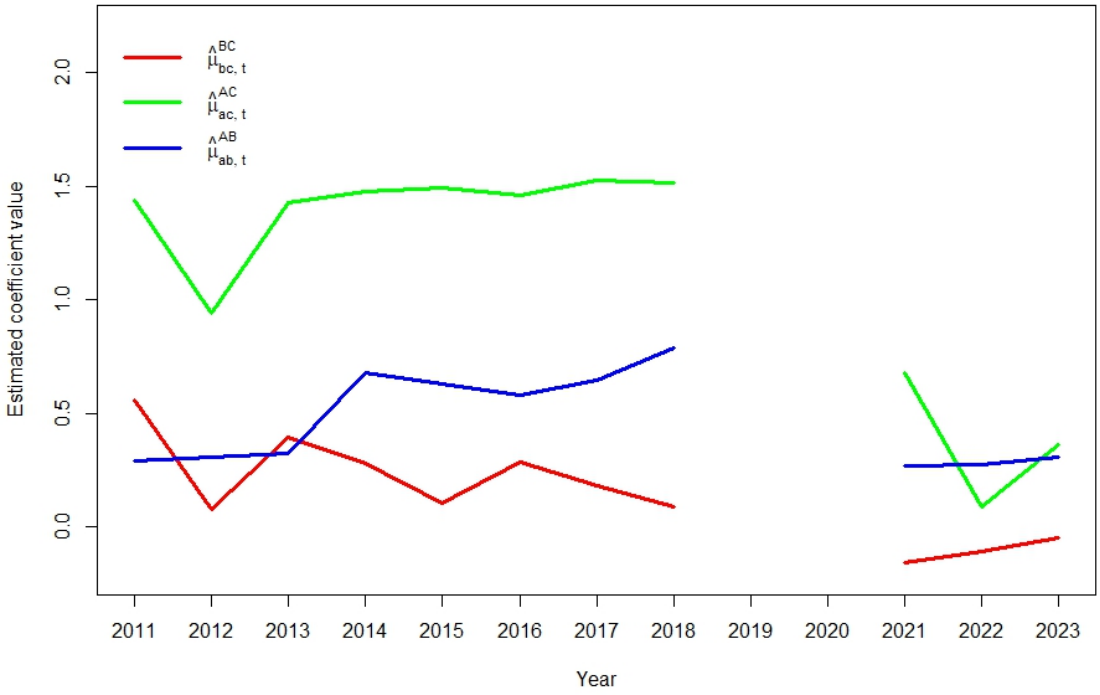
implies that in case of the homeless data the nowcasting model with sample C_y missing gives the best results. This is a bit surprising, because Table 6.3 shows that sample C_y is also the largest sample, which means that its absence should have on average a larger negative impact on the mean absolute difference than the absence of the other sources. However, an explanation of this somewhat paradoxical result can be found in Figure 6.3, which shows that the interaction coefficient $\hat{\mu}^{AB}_{ab,y}$ is more stable than $\hat{\mu}^{AC}_{ac,y}$ and $\hat{\mu}^{BC}_{bc,y}$, and therefore in this example the nowcasting assumption of a stable $\lambda^{AB}_{ab,y}$ is best met when sample C_y is missing, which seems to outweigh the sample size argument.

The first column $\hat{N}^{\text{TSE}}_{(y-1)} - \hat{N}^{\text{TSE}}_{y}$ presents the difference between the current TSE and previous TSE estimate. The mean absolute difference in the last row (4.5) is smaller than two out of three mean absolute differences of the nowcasting models. This can be explained by the relative stability and low volatility of the TSE estimates time series. In case of a less stable or more volatile series, the mean absolute difference will be larger. This implies that in this example of the number of homeless people in The Netherlands, under a different sample delivery order it might be preferable to simply use the lagged time series, but in case of a less stable and more volatile series the nowcasting model may be a better choice.

Finally, to see if the model assumption of stable pairwise-dependencies is reasonable the TSE estimates $\hat{\mu}^{AB}_{ab,y}$, $\hat{\mu}^{AC}_{ac,y}$ and $\hat{\mu}^{BC}_{bc,y}$ over the periods 2011 – 2018 and 2021 – 2023 are presented in Figure 6.3 below.

Figure 6.3 clearly shows three separate time series, which indicates that there is at least some stability in $\mu^{AB}_{ab,y}$, $\mu^{AC}_{ac,y}$ and $\mu^{BC}_{bc,y}$ over time. However, in some years there can be a sudden decrease or increase in the time series, for which we have no immediate

Figure 6.3: Coefficient estimates of $\mu_{ab,v}^{AB}$, $\mu_{ac,y}^{AC}$ and $\mu_{bc,v}^{BC}$ over the periods 2011 – 2018 and 2021 – 2023.



explanation. These large changes correspond to the larger nowcasting errors shown in Table 6.4. Note that in the period 2021 – 2023 the estimate for $\mu_{ac,y}^{AC}$ substantially smaller than in its estimates in the period 2011 – 2018. This can be explained by the fact that sample B_y before 2019 is a different sample than sample B_y after 2019. Before 2019 sample B_y was a sample of homeless people who suffered from drug addictions problems and after 2019 sample B_y was a sample of homeless people of ex-prisoners who received reintegration support.

6.5 Discussion

In this paper we propose to combine dual- and triple-system estimation over two periods by means of the expectation-maximisation algorithm to obtain a preliminary estimate, that we have coined a nowcast estimate. The advantage of this approach is that it allows estimation with two samples, like in DSE, but the independence assumption in DSE is replaced by a more relaxed assumption, which is that the pairwise-dependence of the first two samples is equal to the pairwise-dependence of the first two samples in the previous period. This assumption is more relaxed, because in DSE the independence assumption also implies that the pairwise dependence is equal in two periods, because in DSE the pairwise-dependence should be equal to zero in all periods. This last part of the assumption is not necessary for our proposed nowcast-

ing model. To see if the nowcasting model can be reasonably applied it is therefore advisable, when a sufficiently long time series is available, to check the stability of the interaction parameter estimates.

We applied the TSE nowcasting model to obtain nowcast estimates for the number of homeless people in The Netherlands. The model shows reasonable results in the sense that the nowcast estimates of the expected number of homeless people unique to the missing sample are quite accurate. Furthermore, the nowcasting model estimates are much more similar to the final TSE estimates than the DSE estimates, which indicates that in our example the assumption of stable pairwise-dependency is more realistic than the assumption of pairwise-independence. The accuracy of the nowcasting model is also related to the size of the missing sample. If the largest sample is missing, on average the mean absolute difference between the nowcast and TSE estimate should increase. However, in our case a stable pairwise-dependency was of greater importance than the sample size of the missing sample. Finally, although the TSE nowcasting model provides reasonable results for many periods, we should note that some nowcasting model estimates can be quite inaccurate, for example the nowcasting model estimate $\hat{N}^{\text{NC}}_{ac,y}$ in the years 2021 and 2022, as seen in Figure 6.2. The reason for this inaccuracy was found in the instability of the estimated pairwise-interaction between sample A_y and C_y for those years. Also, because in our example the time series of TSE estimates is reasonably stable, the TSE nowcasting model does not clearly outperform the lagged time series of TSE estimates. Therefore, in cases where the time series of TSE estimates is less stable, the nowcasting model presented in this paper may be more valuable.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, *46*(3), 623–635. Retrieved from https://doi.org/10.2307/2532083

Amstrup, S. C., McDonald, T. L., & Manly, B. F. J. (2005). *Handbook of capture-recapture analysis*. Princeton University Press. Retrieved from https://pubs.usgs.gov/publication/96199

Angelini, E., Banbura, M., & Rünstler, G. (2008). *Estimating and forecasting the Euro area monthly national accounts from a dynamic factor model.* (No. 953). European Central Bank (ECB), Working Paper. Retrieved from https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp953.pdf

Antolin-Diaz, J., Drechsel, T., & Petrella, I. (2021). *Advances in nowcasting economic activity: Secular trends, large shocks and new data.* (No. DP15926). CEPR, Discussion Paper. Retrieved from https://ssrn.com/abstract=3805349

Asimakopoulos, S., Paredes, J., & Warmedinger, T. (2008). *Forecasting fiscal time series using mixed frequency data.* (No. 1550). European Central Bank (ECB), Working Paper. Retrieved from https://www.econstor.eu/bitstream/10419/153983/1/ecbwp1550.pdf

Baffigi, A., Golinelli, R., & Parigi, G. (2004). Bridge models to forecast the Euro area GDP. *Survey Methodology*, *20*, 447–460.

Baffour, B., Brown, J., & Smith, P. (2013). An investigation of triple system estimators in censuses. *Statistical Journal of the IAOS*, *29*, 53–68.

Bailey, N. T. J. (1951). On estimating the size of mobile populations from recapture data. *Biometrika*, *38*(3/4), 293–306.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., ... Groves, R. M. (2010). Research synthesis AAPOR report on online panels. *Public Opin. Q.*, *74*, 711–781.

Bakker, B. F. M., van der Heijden, P. G. M., & Gerritse, S. C. (2017). Estimation of non-registered usual residents in The Netherlands, ultimo September 2010. In D. Bohning, J. Bunge, & P. G. M. van der Heijden (Eds.), *Capture-recapture methods*

for the social and medical sciences (pp. 261–275). CRC Press, Boca Rata. Retrieved from https://doi.org/10.4324/9781315151939-18

Bakker, B. F. M., van Rooijen, J., & van Toor, L. (2014). The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, *30*, 411–424.

Barcellan, R., & Buono, D. (2002). Temporal disaggregation techniques – ECOTRIM Interface (Version 1.01), User Manual, Eurostat. [Computer software manual]. Luxembourg city, Luxembourg. Retrieved from https://ec.europa.eu/eurostat/documents/3859598/9441376/KS-06-18-355-EN.pdf/fce32fc9-966f-4c13-9d20-8ce6ccf079b6

Belser, P., de Cock, M., Mehran, F., & ILO. (2005). *ILO minimum estimate of forced labour in the world*. Geneva, Zwitserland. Retrieved from https://www.ilo.org/media/319681/download

Berkson, J. (1955). Maximum likelihood and minimum χ^2 estimates of the logistic function. *Journal of the American Statistical Association*, *269*(50), 130–162.

Binette, O., & Steorts, R. C. (2022). On the reliability of multiple systems estimation for the quantification of modern slavery. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(2), 640–676.

Bird, S. M., & King, R. (2018). Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Annual review of statistics and its application*, *5*, 95—118.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Springer New York, NY.

Bloem, A. M., Dippelsman, R., & Maehle, N. O. (2001). *Quarterly national accounts manual: Concepts, data sources, and compilation.* Washington, D.C. International Monetary Fund. Retrieved from https://www.imf.org/external/pubs/ft/qna/2000/textbook/index.htm

Box, G. (2013). Box and Jenkins: Time series analysis, forecasting and control. In *A very British affair: Six Britons and the development of time series analysis during the 20th century* (pp. 161–215). London, UK: Palgrave Macmillan UK.

Böhning, D., Rocchetti, I., Maruotti, A., & Holling, H. (2020). Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods. *International Journal of Infectious Diseases*, *97*, 197–201.

Cadwell, B. L., Smith, P. J., & Baughman, A. L. (2005). Methods for capture-recapture analysis when cases lack personal identifiers. *Statistics in Medicine*, *24*(13), 2041–2051.

Cantwell, P. J. (2014). Dual-system estimation. In F. D. Bean & S. K. Brown (Eds.), *Encyclopedia of migration* (pp. 1–5). Springer Netherlands.

Chao, A. (2001). An overview of closed capture–recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, *6*, 158–175.

Chao, A. (2015). Capture-recapture for human populations. In *Wiley statsref: Statistics reference online* (pp. 1–16). John Wiley & Sons, Ltd.

Chao, A., Tsay, P. K., Lin, S. H., & Chao, D. Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine*, *20*, 3123–3157.

Chapman, D. G. (1951). *Some properties of the hypergeometric distribution with applications to zoological sample censuses*. Berkeley, University of California Press. Retrieved from https://babel.hathitrust.org/cgi/pt?id=wu.89045844248&view=1up&seq=3

Chapman, D. G. (1952). Inverse, multiple and sequential sample censuses. *Biometrics*, *8*(4), 286–306.

Chapman, D. G. (1954). The estimation of biological populations. *The Annals of Mathematical Statistics*, *25*(1), 1–15. Retrieved from https://www.jstor.org/stable/2236510

Chatterjee, K., & Mukherjee, D. (2018). A new integrated likelihood for estimating population size in dependent dual-record system. *The Canadian Journal of Statistics*, *46*(4), 577–592.

Chen, Q., & Giles, D. (2011). Finite-sample properties of the maximum likelihood estimator for the Poisson regression model with random covariates. *Communications in Statistics—Theory and Methods*, *40*, 1000–1014.

Chen, Z., & Kuo, L. (2001). A note on the estimation of the multinomial logit model with random effects. *The American Statistician*, *55*(2), 89–95. Retrieved from http://www.jstor.org/stable/2685993

Chow, G., & Lin, A. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics*, *53*(4), 372–375.

Cochran, W. G. (1978). Laplace's ratio estimator. In H. DAVID (Ed.), *Contributions to survey sampling and applied statistics* (pp. 3–10). Academic Press.

Cordeiro, G. M., & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(3), 629–643. Retrieved from https://www.jstor.org/stable/2345592

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics*, *45*(2), 395–413.

Cormack, R. M., & Jupp, P. E. (1991). Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrica*, *78*(4), 911–916.

Coumans, M. A., Cruyff, M., van der Heijden, P. G. M., Wolf, J., & Schmeets, H. (2017). Estimating homelessness in The Netherlands using a capture-recapture approach. *Social Indicators Research*, *130*(1), 89–212.

Cramer, H. (1922). *Mathematical methods of statistics*. Princeton University Press, London. Retrieved from https://archive.org/details/in.ernet.dli.2015.149716/page/n515/mode/2up

Daalmans, J. A. (2018). Special issue article: Benchmarking, temporal disaggregation, and reconciliation of systems of time series. *Statistica Neerlandica*, *72*(4), 406–420.

Dagum, E. B., & Cholette, P. A. (1975). *Benchmarking, temporal distribution, and reconciliation methods for time series. Part of the book series: Lecture Notes in Statistics.* (Vol. 186). Springer Science & Business Media.

Darroch, J. N. (1958). The multiple-recapture census: I. Estimation of a closed population. *Biometrika*, *45*(3/4), 343–359.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., & Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, *88*(423), 1137–1148.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. Retrieved 2024-03-21, from http://www.jstor.org/stable/2984875

Denton, F. T. (1971). Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, *66*(333), 99–102.

de Wolf, P., van der Laan, J., & Zult, D. B. (2019). Connecting correction methods for linkage error in capture-recapture. *Journal of Official Statistics*, *35*(3), 577–597. Retrieved from https://doi.org/10.2478/jos-2019-0024

Di Consiglio, L., & Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, *31*(3), 415–429.

Di Consiglio, L., & Tuoto, T. (2018). Population size estimation and linkage errors: the multiple lists case. *Journal of Official Statistics*, *34*, 889–908.

Ding, Y., & Fienberg, S. E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology*, *20*(2), 149–158. Retrieved from https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199400214422

Doornik, J. A. (2009). An object-oriented matrix programming language Ox 6 [Computer software manual]. London: Timberlake Consultants Press.

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of economics and statistics*, *94*(4), 1014–1024. Retrieved from http://www.jstor.org/stable/23355337

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods, second edition*.

Eurostat. (2008). *Nace rev. 2, statistical classification of economic activites in the European Community*. Retrieved from https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF

Eurostat. (2017). *Handbook on rapid estimates, 2017 edition.* Retrieved from https://ec.europa.eu/eurostat/documents/3859598/8555708/KS-GQ-17-008-EN-N.pdf/7f40c70d-0a44-4459-b5b3-72894e13ca6d?t=1513758176000

Eurostat. (2018). *Ess guidelines on temporal disaggregation, benchmarking, and reconciliation, 2018 edition.* Retrieved from https://ec.europa.eu/eurostat/documents/3859598/9441376/KS-06-18-355-EN.pdf/fce32fc9-966f-4c13-9d20-8ce6ccf079b6

Evans, M. A., & Bonett, D. G. (1994). Bias reduction for multiple-recapture estimators of closed population size. *Biometrics*, *50*(2), 388–395.

Evans, M. A., Bonett, D. G., & McDonald, L. L. (1994). A general theory for modeling capture-recapture data from a closed population. *Biometrics*, *50*(2), 396–405.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Fernández, R. (1981). A methodological note on the estimation of time series. *The Review of Economics and Statistics*, *63*(3), 471–476.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, *59*(3), 591–603.

Fienberg, S. E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, *18*, 143–154. Retrieved from https://www150.statcan.gc.ca/n1/pub/12-001-x/1992001/article/14494-eng.pdf

Fienberg, S. E., Johnson, M. S., & Junker, B. W. (2002). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *162*(3), 383–405.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38.

Frome, E. L., Kutner, M. H., & Beauchamp, J. J. (1973). Regression analysis of Poisson-distributed data. *Journal of the American Statistical Association*, *68*(344), 935–940.

Gerritse, S. C., Bakker, B. F. M., de Wolf, P., & van der Heijden, P. G. M. (2016a). *Under coverage of the population register in The Netherlands.* Discussion paper 2016-02 (Centraal Bureau voor de Statistiek, Den Haag/Heerlen). Retrieved from https://dspace.library.uu.nl/bitstream/handle/1874/356071/register.pdf?sequence=1

Gerritse, S. C., Bakker, B. F. M., Zult, D. B., & van der Heijden, P. G. M. (2016b). *The effects of imperfect linkage and erroneous captures on the population size estimator, chapter 3 of phd thesis* (Doctoral dissertation). Retrieved from https://www.cbs.nl/en-gb/background/2017/39/impact-of-linkage-errors-and-erroneous-captures

Gerritse, S. C., van der Heijden, P. G. M., & Bakker, B. F. M. (2015). Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of Official Statistics*, *80*(3), 357–379.

Ghysels, E., Kvedaras, V., & Zemlys, V. (2016). Mixed Frequency Data Sampling regression models: The R package midasr. *Journal of Statistical Software*, *72*(4), 1—35.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). *The MIDAS touch: Mixed data sampling regression models.* Retrieved from https://escholarship.org/uc/item/9mf223rs

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*(1), 53–90.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676.

Gill, G. V., Ismail, A. A., Beeching, N. J., Macfarlane, S. B., & Bellis, M. A. (2003). Hidden diabetes in the UK: use of capture–recapture methods to estimate total prevalence of diabetes mellitus in an urban population. *Journal of the Royal Society of Medicine*, *96*(7), 328—332.

Hald, A. H. (1952). *Statistical theory with engineering applications*. John Wiley & Sons, Inc. Retrieved from https://archive.org/details/statisticaltheor0000ahal/mode/2up?view=theater

Hald, A. H. (1975). *A history of probability and statistics and their applications before 1750*. New York: Wiley. Retrieved from https://archive.org/details/historyofprobabi0000hald

Hammond, C., van der Heijden, P. G. M., & Smith, P. A. (2024). Generating contingency tables with fixed marginal probabilities and dependence structures described by loglinear models. *Journal of Statistical Computation and Simulation*, *94*(12), 2797–2812.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques* (Second ed.). Springer. Retrieved from https://link.springer.com/book/10.1007/0-387-69505-2

Hogan, H., Cantwell, P., Devine, J., Mule, V., & Velkoff, V. (2013). Quality and the 2010 census. *Population Research and Policy Review*, *32*, 637-–662.

Hook, E. B., & Regal, R. R. (1995). Capture-Recapture Methods in Epidemiology: Methods and Limitations. *Epidemiologic Reviews*, *17*(2), 243–264.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3), 1–22.

ILO. (2018, October). Guidelines concerning the measurement of forced labour. Geneva, Zwitserland. Retrieved from https://www.ilo.org/media/209456/download

International Working Group for Disease Monitoring and Forecasting. (1995a). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology*, *142*(10), 1047–1058.

International Working Group for Disease Monitoring and Forecasting. (1995b). Capture-recapture and multiple-record systems estimation II: Applications in human diseases. *American Journal of Epidemiology*, *142*(10), 1059–1068.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420. Retrieved from http://www.jstor.org/stable/2289924

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, *52*(1/2), 225–247. Retrieved from http://www.jstor.org/stable/2333826

Koopman, S. J., Shephard, N., & Doornik, J. A. (2008). Ssfpack 3.0: Statistical algorithms for models in state space form. [Computer software manual]. London: Timberlake Consultants Press.

Kosmidis, I. (2007). *Bias reduction in exponential family nonlinear models* (Doctoral dissertation, The University of Warwick). Retrieved from https://www.ikosmidis.com/files/ikosmidis_thesis.pdf

Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *WIREs Comput Stat*, *6*(3), 185–196.

Kosmidis, I., & Firth, D. (2011). Multinomial logit bias reduction via the Poisson log-linear model. *Biometrika*, *98*(3), 755–759. Retrieved from https://www.jstor.org/stable/23076146

Kosmidis, I., & Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometirka*, *108*, 71–82.

Kosmidis, I., & Kenne Pagui, E. C. (2023). brglm2: Bias reduction in generalized linear models [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/brglm2/brglm2.pdf

Kosmidis, I., Kenne Pagui, E. C., & Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing*, *30*, 43-–59.

Lincoln, F. C. (1930). *Calculating waterfowl abundance on the basis of banding returns* (Vol. 118). United States Department of Agriculture.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, *5*, i-iii+v-ix+1–163. Retrieved from http://www.jstor.org/stable/4153184

Litterman, R. B. (1983). A random walk, Markov model for the distribution of time series. *Journal of Business and Economic Statistics*, *1*(2), 169–173.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables* (Vol. 7). SAGE Publications, Inc. Retrieved from https://us.sagepub.com/en-us/nam/regression-models-for-categorical-and-limited-dependent-variables/book6071

Lum, K., Price, M. E., & Banks, D. (2013). Applications of multiple systems estimation in human rights research. *The American Statistician*, *67*(4), 191–200. Retrieved from http://www.jstor.org/stable/24591478

Manrique-Vallier, D., Price, M. E., & Gohdes, A. (2013). Multiple systems estimation techniques for estimating casualties in armed conflicts. In *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict*. Oxford University Press.

McClintock, B. T., Conn, P. B., Alonso, R. S., & Crooks, K. R. (2013). Integrated modeling of bilateral photo-identification data in mark–recapture analyses. *Ecology*, *94*(7), 1464–1471.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Second ed.). London: Chapman and Hall. Retrieved from https://doi.org/10.1201/9780203753736

McLeod, P., Heasman, D., & Forbes, I. (2011). *Simulated data for the on the job training, ESSnet DI*. Retrieved 2015, from https://ec.europa.eu/eurostat/cros/content/job-training_en

Menkens, G. E. J., & Anderson, S. H. (1988). Estimation of small-mammal population size. *Ecology*, *69*(6), 1952–1959.

Miller, D. M. (1984). Reducing transformation bias in curve fitting. *The American Statistician*, *38*(2), 124–126. Retrieved from https://doi.org/10.2307/2683247

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society.*, *26*(9), 394–395. Retrieved from https://doi.org/10.1090/S0002-9904-1920-03322-7

Muneza, A. B., Linden, D. W., Montgomery, R. A., Dickman, A. J., Gary, J. R., Macdonald, D. W., & Fennessy, J. T. (2017). Examining disease prevalence for species of conservation concern using non-invasive spatial capture–recapture techniques. *Journal of Applied Ecology*, *54*, 709—717.

Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, *62*, 3–135. Retrieved from https://www.jstor.org/stable/3830650

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society.*, *51*(3), 406–413.

Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, *6*, 5–84. Retrieved from https://archive.org/details/reportofdanishbi06dans/page/n1/mode/2up

Plackett, R. L. (1981). *The analysis of categorical data* (Second ed.). New York: Macmillan. Retrieved from https://catalogue.nla.gov.au/catalog/171489

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rainey, C., & McCaskey, K. (2021). Estimating logit models with small samples. *Political Science Research and Methods*, *9*(3), 549–564.

Rivest, L. (2022). Rcapture: Loglinear models for capture-recapture experiments [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/Rcapture/Rcapture.pdf

Rivest, L., & Lévesque, T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *The Canadian Journal of Statistics*, *29*(4), 555–572.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, *130*(1), 142–152. Retrieved from https://www.jstor.org/stable/2239906

Sax, C., & Steiner, P. (2013). *tempdisagg: Methods for temporal disaggregation and interpolation of time series.* Retrieved from https://cran.r-project.org/web/packages/tempdisagg/tempdisagg.pdf

Schnabel, Z. E. (1938). The estimation of total fish population of a lake. *The American Mathematical Monthly.*, *45*(6), 348–352.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika*, *52*(1/2), 249–259. Retrieved from http://www.jstor.org/stable/2333827

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters* (Second ed.). London: Griffin. Retrieved from https://archive.org/details/estimationofanim0000sebe/page/n5/mode/2up

Sekar, C. C., & Deming, E. W. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, *44*(245), 101–115. Retrieved from http://www.jstor.org/stable/2280353

Silverman, B. W. (2020). Multiple-systems analysis for the quantification of modern slavery: Classical and Bayesian approaches. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*, 691-–736.

Silverman, B. W., Chan, L., & Vincent, K. (2023). Bootstrapping multiple systems estimates to account for model selection. *Statistics and Computing*, *34*(44), 156–177.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48. Retrieved from https://doi.org/10.2307/1912017

Sowden, B. R. (1972). On the first-order bias of parameter estimates in a quantal response model under alternative estimation procedures. *Biometrika*, *59*(3), 573–579. Retrieved from http://www.jstor.org/stable/2334808

Statistic Netherlands. (2016). *Usual residence population definition: Feasibility study The Netherlands.* Retrieved from https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende-onderzoeksomschrijvingen/usual-residence-population-definition

Stephan, F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate. *The Annals of Mathematical Statistics, Ann. Math. Statist., 16*, 50-–61.

Stock, J. H., & Watson, M. W. (1980). Vector autoregressions. *The journal of economic perspectives*, *15*(4), 101–115.

Tilling, K. (2001). Capture-recapture methods—useful or misleading? *International Journal of Epidemiology*, *30*(1), 12–14.

Tilling, K., & Sterne, J. A. (1999). Capture-recapture models including covariate effects. *American journal of epidemiology*, *149*(4), 392–400.

UNODC. (2022). *Monitoring human trafficking prevalence through multiple systems estimation.* Retrieved from https://www.unodc.org/documents/data-and-analysis/tip/2022/MSE_TIP_UNODC_ENG.pdf

van der Heijden, P. G. M., Cruyff, M., Smith, P. A., Bycroft, C., Graham, P., & Matheson-Dunning, N. (2021). Multiple system estimation using covariates having missing values and measurement error: Estimating the size of the Māori population in New Zealand. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(1), 156–177.

van der Heijden, P. G. M., Whittaker, J., Cruyff, M., Bakker, B. F. M., & van der Vliet, R. (2012). People born in the Middle East but residing in The Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics*, *6*(3), 831–852.

White, S. R., Bird, S. M., & Grieve, R. (2014). Review of methodological issues in cost-effectiveness analyses relating to injecting drug users, and case-study illustrations. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *177*(3), 625–642. Retrieved from http://www.jstor.org/stable/43965417

Winkler, W. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Section on Survey Research Methods*, 667–671. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2000/adrm/rr2000-05.pdf

Wittes, J. T. (1972). 331. Note: On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate. *Biometrics*, *28*(2), 592–597. Retrieved from http://www.jstor.org/stable/2556173

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, *81*, 338–346.

Yauck, M., Rivest, L., & Rothman, G. (2019). Capture-recapture methods for data on the activation of applications on mobile phones. *Journal of the American Statistical Association*, *114*(525), 105–114.

Zaslavsky, A. M., & Wolfgang, G. S. (1993). Triple-system modeling of census, post-enumeration survey, and administrative-list data. *Journal of Business & Economic Statistics*, *11*(3), 279–288.

Zhang, L. C. (2019). A note on dual system population size estimator. *Journal of Official Statistics*, *35*(1), 279–283.

Zult, D. B., de Wolf, P., Bakker, B. F. M., & van der Heijden, P. G. M. (2021). A general framework for multiple-recapture estimation that incorporates linkage error correction. *Journal of Official Statistics*, *37*(3), 699–718.

Zult, D. B., Krieg, S., Schouten, B., Ouwehand, P., & van den Brakel, J. (2020). *From quarterly to monthly turnover figures using nowcasting.* Statistics Netherlands, Discussion paper. Retrieved from https://www.cbs.nl/-/media/_pdf/2020/14/nowcasting-fribs-24-maart.pdf

Zult, D. B., Krieg, S., Schouten, B., Ouwehand, P., & van den Brakel, J. (2023). From quarterly to monthly turnover figures using nowcasting methods. *Journal of Official Statistics*, *39*(2), 253–273.

Zult, D. B., van der Heijden, P. G. M., & Bakker, B. F. M. (2023). Bias correction in multiple-systems estimation. *arXiv*. Retrieved from https://doi.org/10.48550/arXiv.2311.01297

Zult, D. B., van der Heijden, P. G. M., & Bakker, B. F. M. (2024). Nowcasting in triple-system estimation. *arXiv*. Retrieved from https://arxiv.org/abs/2406.17637

Zwane, E. N., & van der Heijden, P. G. M. (2007). Analysing capture–recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations. *Statistics in Medicine*, *26*, 1069–1089.

Zwane, E. N., van der Pal-de Bruin, K., & van der Heijden, P. G. M. (2004). The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in medicine*, *23*, 2267–81.

Acknowledgements

Writing a dissertation has been a long journey that already started in another period of my life, at a different place and on a different topic, in a time when apparently I was not ready for such a commitment. I will not discuss the details of this episode, but I think it is important to mention that this dissertation is my second try, because this makes the patience, faith and support that I received from my two supervisors Bart Bakker and Peter van der Heijden, even more admirable. I am very grateful for their support and without it, this dissertation would not have seen the light of day. Their patience, faith and support was further tested by the substantial time it took to write this dissertation. There were periods in which progress was slow or even non-existent, sometimes because some of our ideas turned out to be unfit for the general topic of this dissertation, the desired data was unavailable, I was busy with other work at Statistics Netherlands, or my family with my young children Olav and Nova deserved additional attention, especially during the COVID-19 pandemic.

First and foremost I want to thank my earliest supervisor Bart, who first accepted me as an employee at the Methodology department of Statistics Netherlands, a great place with friendly, smart and knowledgeable colleagues, which allowed me to further develop my interest and skills on statistical methods within various fields. Bart was also open to the idea of acting as my dissertation supervisor, and he accepted me as his PhD-student, a great compliment. As a supervisor he gave me the confidence to believe that there is a place in science for me and my approach to science, including my shortcomings. Bart continued to support me after he suffered from serious health issues and after his retirement, for which I will always be grateful. I can imagine that there must have been moments in which he regretted our collaboration, but I can say that this regret is not mutual :).

Where Bart played a critical role in the beginning of this project, Peter played a pivotal role in finishing it. In the final period Peter and I had regular contact in shared documents and online sessions, in which Peter patiently continued to give me helpful and instructive feedback on my ideas, analyses and writings. These sessions helped me to give structure to the ideas presented in this dissertation and to present and do the analyses in a more systematic way. I think I learned a lot from these sessions. I am very grateful for Peter's commitment to our sessions, especially because I know I can be (a bit too) stubborn which did not make it easier for him. Therefore, Peter, thank you again for your enduring patience and for all of your lessons, you can rest assured that I will remember and use them in the future!

Although he was not involved with the content of this dissertation, also my current manager at the Methodology department, Reinoud Stoel, deserves my gratitude. The

last few years Reinoud has been my manager at the Methodology department and he supported me by both pushing me in a positive way and by supporting opportunities for research that were beneficial for both Statistics Netherlands and this dissertation. Reinoud, your support was of crucial importance, without it you would not be reading this page with these words of appreciation.

There are many colleagues who contributed to the content of this dissertation in some way or another, but I would like to mention a few by name. First, Moniek Coumans, with whom I had many pleasant and fruitful discussions on the topic of estimating the number of homeless people in The Netherlands. Moniek helped me to understand the data on homeless people and she provided me with data whenever I had a special data request. Her help was crucial for the analyses presented in Chapter 2 and Chapter 6.

Furthermore, I should mention Peter-Paul de Wolf and Jan van der Laan, whose insights and contributions to Chapter 3 and 4 on linkage error corrections were vital. With them, I had many thoughtful discussions on the mathematical foundations of multiple systems estimation in general, and how the method can be combined with probabilistic record linkage models. Peter-Paul, Jan and I worked closely on this topic at the beginning of this dissertation project, and I have pleasant memories of our trip together to ISTAT. On this trip, I also had the chance to meet a group of multiple systems estimation experts, such as Tiziana Tuoto and Li-Chun Zhang, which was very helpful and inspiring.

Chapter 5, which is the only topic that is not directly related to multiple systems estimation, but only indirectly trough time series analysis, gave me the pleasant opportunity to collaborate with some of the best time series analysis experts that work at Statistics Netherlands, both from The Hague and Heerlen. It gave me the chance to work together with Sabine Krieg, Pim Ouwehand and Jan van den Brakel, who each have a long history in academic research on time series analysis. The COVID-19 pandemic lengthened and deepened our collaboration, because it forced us to think about the performance of time series models before, during and after a crisis. Working with Sabine, Pim and Jan was a pleasure and deepened my knowledge on time series models, for which I want to thank them. With respect to Chapter 5, I should also mention the enormous help I received from Bart Klein, with whom I discussed the nowcasting model results in great detail. Bart also did a great job improving the quality of the data that was used for Chapter 5, which greatly improved the quality of the analysis.

Beside everybody who somehow contributed in some way to specific parts of this dissertation, I want to thank each of my excellent colleagues at the Methodology department in The Hague, who were always open and available to help me whenever I consulted any of them. In particular I would like to thank Sander Scholtus and Jeroen Pannekoek, who read some of the parts in this dissertation carefully and suggested some valuable improvements.

Beside my colleagues at Statistics Netherlands, I am also indebted to the five renowned experts who accepted to become part of my doctoral examining committee. Dear Prof. dr. Katrijn van Deun, Dr. Peter Lugtig, Prof. dr. Daniel Oberski, Prof.

dr. Barry Schouten and Prof. Paul Smith, some of us have met before but I want to thank each of you equally for accepting this commitment and responsibility. I hope that each of you enjoyed reading this dissertation and I am looking forward to discussing it with you during the defense.

I also want to pay a tribute to all the anonymous reviewers in general and in particular those who helped me with their instructive comments, remarks, critiques and suggestions on the chapters in this dissertation. I don't know who or where you are, but I can say that I learned a lot from you and you helped me to improve the quality of this dissertation, for which I thank you sincerely.

Last but not least I want to thank my family and friends for all the love, joy and friendship they gave me. Of course, this transcends the writing of a dissertation, but without it this text would not have been written. I first and foremost want to thank my better half Sylvia for her everlasting support and confidence. Without our love, companionship and her care this dissertation would not exist, but in that case this would have been the least of my worries. I want to thank my loving and supporting parents, who gave me the freedom to choose my own future, and my children Olav and Nova, for demanding my attention without caring if my mind is somewhere else. I also should not forget Sylvia's parents Ineke and Johan, whom I love dearly and who, just like my own parents, gave me the time to write this dissertation by taking care of Olav and Nova so many times and with so much love.

And finally my friends, who unwillingly taught me that life can be cruel, or to paraphrase Bilbo Baggins, but without any irony: I don't see half of you half as many times as I should like; and I see less than half of you half as much as you deserve. I have fond memories of each of you, some of these memories go back to high-, or even primary-, school in Hoorn, and many memories go back to more recent times often in Amsterdam, or from the verge of a chess-board. I hope that finishing this dissertation, and with Olav en Nova a bit older, it is time for making more of such fond memories.