An application of population size estimation to official statistics.

Sensitivity of model assumptions and the effect of implied coverage.

Susanna Charlotte Gerritse

.

Cover Illustration by
Lucky Cat Creations, 2016

Printed by Gildeprint,
ISBN: 978-94-6233-323-9

An application of population size estimation to official statistics.
Sensitivity of model assumptions and the effect of implied coverage.

Een toepassing van populatieschattingen voor de officiele statistiek.
Gevoeligheid van model aannames en het effect van implied coverage
(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. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar ter verdediging op vrijdag 8 juli 2016 des ochtends te 10.30 uur door

Susanna Charlotte Gerritse
geboren op woensdag 22 juni 1988 te Amsterdam

Promotors: Prof. Dr. P.G.M. van der Heijden
Prof. Dr. B.F.M. Bakker

To Mama, my ’Pappie’ and Frank.

And my love, Merlin.

”A thousand signs declined
That traveled through light
Translate this mystery
That covered my eyes
Accept approaching fear
And courage appears
Death is a certainty
It’s growing near

Letting go is fateful

...

I’ll come on home
I’m in the light of day
Questions were answered
A new life’s arms are extending
The final page has turned, sending the letter
Come on home
And I’ll sing you the song that has painted your canvas of life”

Epica - Canvas of life

Foreword

The essence of all beautiful art is gratitude.
Friedrich Nietzche

I'd like to start with thanking the methodology department at Statistics Netherlands for housing me and acknowledging me as one of their own. In particular I would like to name those CBS colleagues that were extra special and for whom I hold a great deal of respect: Jacobiene van der Hoeven, Sander Scholtus, Rik van der Vliet, Ton de Waal, Daan Zult. Eric Schulte Nordholt, for all the advise, laughs and lunches. Barteld Braaksma for always motivating me and introducing me to new amazing people, you rock! I am truly grateful for all you did for me and believing in me. You are a mentor and a sponsor all in one and I could not think of a better person for those jobs. Peter-Paul de Wolf, thank you for all the help and laughs, you were the best (unfortunately, un-official) daily supervisor. I miss our monday-morning talks at the coffee machine. A special thanks goes to official statistics colleague from Ireland: John Dunne, I hold a great deal of respect for you and I am grateful to all you have done.

A big thanks goes out to the Methods and Statistics Department of Utrecht University. To everyone: Thank you for the amazing time. A special thanks go out to: Manon Bouman, Rens van der Schoot, Noemi Schuurman, Vera Toepoel. And all my roommates: Sander van Schie, Marielle Zondervan-Zwijnenburg, Sharon Klaassen, Sanne Smid. A special acknowledgement goes to Frank Bais for all the fun, your amazing wit and our lunches. I will remember to write something up! But especially to Anouck Kluytmans, for everything, love you girl!

At the UU I found an outlet for my curiosity and activism in the form of Prout and PNN. I would like to thank all the Prout members for the amazing and fun time I had, especially Sophie van Uijen and Jeroen Goudsmit. To all former, current and future PNN members: I want to thank you! PNN deserves way more credit than I can give here in this dedication, but trust me when I say I never met a bunch of amazing and fun people with such a passion and enthusiasm all bundled for one greater cause.

Of all the colleagues, there are two men I have to thank in particular: Professor Peter G. M. van der Heijden and Professor Bart F. M. Bakker. These men have challenged me to be a better researcher and also a better person, and I look back on them fondly. Peter, I've come to respect you as a researcher and a caring and funny person. I did not know you as a researcher before the PhD, but during I found out so many people knew you and respected you, it is an honor to tell people I graduated under you. Bart, I value everything you have done for me and what I have learned from you, even though I may have never been able to actually put that to words. I can honestly tell you that I never had so much respect for a person. You are a kindhearted man and a great researcher. I hope to continue to have a laugh or two with the both of you in the future. Thank you, both.

Then there are the people I would not have ever even been here without. My parents, Louise and Henk. I love you so much, and I thank you for everything you have ever taught me. It is with pride I dedicate this PhD to the both of you. My grandparents, even though I have to miss half of them. My little brother Frank (enunciating the little, given he is 5 years younger, yet 15 cm taller), we have laughed and we have fought, and without you life would be dull and empty. I am so proud of you, and I hope I made you proud of your sister. My in laws, Klaas and Sylvia, you gave me my amazing life partner, and you have supported me throughout it all. My amazing friends Gabrielle and Naomi Verwer, Lisette Rodenhuis-Van Mourik, Danielle Overdijk, and Samantha Bakx, for always supporting me. Lisette, thank you for the cover art. But mostly for always being there and being the best friend ever! A special thanks go to my paranimfs: Anouck, I am sorry this job as a paranimf is not more exciting (wink), especially since from the beginning of my PhD I knew you were going to be my paranimf. Gabrielle, we have been friends for a long time, and I hope to stay friends for even longer! I love you guys.

Ruben, it may have been your job, but you gave me back my life. I am forever grateful. I think your voice and your advise are cemented into my brain, never to be forgotten. You gave me back my trust.

Last, but most definitely not least: Merlin von Freytag Drabbe. You have been with me for 10 years now. I cannot put to words what I would have done without you. You always tell me: "the same as you would with me". But that is not true. I can not imagine my life without you anymore, and I would not have it any other way, I love you. Je bent zo leuk he!

Contents

List of Figures

List of Tables

1 Introduction

Susanna C Gerritse

Statistics Netherlands

CONTENTS

1.1 Introduction

Official Statistics bureaus are periodically asked to give an estimate of their country's population, which can be defined by the number of usual residents. According to EU Regulation No 1260/2013, usual residence is defined by the place where a person normally spends the daily period of rest. Then, a person is considered a usual resident when they have lived in the Netherlands for longer than a year, or if they have the intention to reside for longer than a year. For the Dutch Census, Statistics Netherlands makes use of the Population Register (PR). By Dutch law, every individual that is residing, or planning to reside, in the Netherlands for longer than four months, has to register in the PR. However, for numerous reasons, immigrants that have taken residence in the Netherlands may not register and become undocumented immigrants. Given that the Population Register only consists of individuals that actually have registered themselves, the PR alone is not sufficient to estimate the number of usual residents, and has an undercoverage considering the number of Dutch usual residents.

One commonly used method to estimate population sizes is the capture-recapture methodology. First the PR is linked to two other registers. Then capture-recapture methodology using a covariate that denotes residence duration can be used to estimate the number of usual residents missed by all three registers. However, for the valid use of capture-recapture methodology, a set of assumptions has to be met. Additionally, practical issues such as missing data may occur. Such practical issue have to be resolved before one can estimate the number of Dutch usual residents via capture-recapture methodology. For that purpose there are two central questions in this thesis: 1) what is the effect of violated assumptions and missing data on the robustness of population size estimation via capture-recapture methodology, and 2) how can the information gained in 1) be used to achieve a trustworthy estimate of the under coverage of usual residents in the Population Register in the Netherlands?

To answer the first question in this thesis, research has been conducted into the robustness of population size estimation via capture-recapture methodology when three specific assumptions have been violated. These assumptions are 1) independence of the inclusion probabilities of the registers, 2) no erroneous captures in the registers, such that all units in the registers belong to the population and 3) perfect linkage of the units in the used registers. For the independence assumption, this research also investigated the robustness for independence conditional on fully and partially observed covariates. Additionally research has been conducted into the effect missing data have on the population size estimation, and most notably how different methods of handling missing data differ in their effect on the resulting population size estimate.

To assess the effect of violated assumptions on population size estimation, capture-recapture methodology has been conducted on two different nationality groups. One group considered only individuals with a Polish nationality, and the other group considered individuals with an Afghan, Iraqi and Iranian nationality. These nationality groups differed in implied coverage. The idea that implied coverage is important for population size estimation has been considered before in previous research, although it has not been called such, but rather implied dependence Brown et al. (2006), and has been studied more thoroughly in this thesis. Implied coverage can be explained by considering two registers, register 1 and register 2. Register 1 is the register containing the highest number of individuals, such that of the two registers, register 1 will have the highest coverage of the population. Implied coverage indicates the size of the coverage register 1 has of the population, given register 2. Thus implied coverage describes the relative size of new cases provided by register 2. Implied coverage plays an important role in this thesis given that it cannot be ascertained from the data whether assumptions are violated, but implied coverage can. It turns out that it can be assessed whether violated assumptions will result in a large biasing effect or not and whether conclusions have to be drawn cautiously.

To answer the second central question in this thesis, the results obtained in answering the first question have been used to conduct research into the under coverage of the PR of the Netherlands. The steps that are taken to estimate the undercoverage of the PR using capture-recapture methodology are discussed in the last chapter of this thesis. This resulted in an estimate of the undercoverage of the population.

This chapter starts with discussing some theoretical background of capture-recapture methodology and the assumptions. This information is needed to answer the first main research question. Then, this chapter will discuss some practical background information, which comprises mostly of information on Statistics Netherlands and the data used. This chapter ends with overview of the contribution of this thesis.

1.2 Theoretical background to this thesis

1.2.1 Short overview of capture-recapture methodology

Capture-recapture methodology is a common method to estimate hard-to-reach populations, and goes by many names, such as the Petersen method, the Lincoln Index, mark-recapture and dual and multiple system estimation. Currently capture-recapture methodology is used to estimate a variety of hard-to-reach populations. A few examples are: estimating rare diseases such as neural tube defects (Zwane and Van der Heijden, 2008), undocumented immigrants (Van der Heijden et al., 2011), hard-drug users in the Netherlands (Cruts and Van Laar, 2010), children under 15 years of age injured in motor vehicle accidents (Jarvis et al., 2000) or the prevalence of individuals with HIV in France (Hraud-Bousquet et al., 2012).

One of the first applications of capture-recapture estimation of human populations is by Laplace in 1786. Sekar and Deming (1949) estimated registered birth and death rates and compared these to actual birth and death rates. A mathematical framework for capture-recapture methodology has been proposed by Darroch (1968), for a multiple recapture census of animals. He proposed a sequence of k samples, where every member of the ith sample is uniquely labeled. This labeling is done via marking or tagging, after which the animal is returned to the population. Fienberg (1972) and Cormack (1989) showed how the multinomial multiple recapture census can be reparametrized into a loglinear model.

Some reviews and books on capture-recapture methodology have been published. An overview of the design, inference and interpretation of capture-recapture experiments for animal populations can be found in different reviews such as Otis et al. (1978) and Pollock et al. (1990), or books such as Seber (1982).

There are also books and overviews of capture-recapture estimation of human populations. A synthesis on capture-recapture methodology was published by Cormack (1968), followed by a book focussing on the technical aspects of the method (Cormack, 1979). An influential book on multivariate analysis of categorical data, considering also incomplete tables, is Bishop et al. (1975). Examples of reviews of capture-recapture methodology and its application to human populations are, amongst other (International Working Group for Disease Monitoring and Forecasting, 1995a, 1995b) and Fienberg (1992) on the application to the US Census. Researchers Hook and Regal have published a couple of full length articles on capture-recapture methodology, the theory and a few applications amongst others to multiple lists (up to 14 lists) and one of the earliest investigations to variable catchability (Hook and Regal, 1992, 1993, 1995, 1997, 2000).

1.2.2 Two register capture-recapture

The simplest population size estimation model makes use of two registers, register 1 and register 2. First the individuals have to be correctly identified as being registered, or not registered in register 1 and register 2, resulting in a 2 by 2 contingency table. There is one zero count for the individuals that are part of the population but were not registered in either register (Bishop et al., 1975). By using capture-recapture methodology we can estimate this zero count.

Let variables A and B respectively denote inclusion in registers 1 and 2. Let the levels of A be indexed by i ($i = 0,\ 1$) where $i = 0$ stands for "not included in register 1", and $i = 1$, stands for "included in register 1". Similarly, let the levels of B be indexed by j ($j = 0, 1$). Expected values are denoted by m_{ij}. Observed values are denoted by n_{ij} with $n_{00} = 0$, because there are no observations for the cases that belong to the population but were not present in either of the registers.

TABLE 1.1
Expected values of being present in register 1 and register 2

	A		
B	1	0	Total
1	m_{11}	m_{10}	m_{1+}
0	m_{01}	m_{00}	m_{0+}
Total	m_{+1}	m_{+0}	m_{++}

Table 3.1 shows the expected values. If expected value m_{00} would have been observed by n_{00}, the saturated loglinear model would be

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \tag{1.1}$$

where we use the identifying restrictions $\lambda_0^A = \lambda_0^B = \lambda_{00}^{AB} = \lambda_{10}^{AB} = \lambda_{01}^{AB} = 0$. Parameter λ_{11}^{AB} can be used to estimate the dependence of the inclusion probabilities of the registers. We use the notation of Bishop et al. (1975) for hierarchical loglinear models, where the saturated model is denoted [AB].

However we only have observed counts n_{11}, n_{10} and n_{01}, given that n_{00} is a structural zero and has to be estimated. Since there are three counts, only three parameters can be estimated, and we have to assume independence. Thus the saturated log-linear model is:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B. \tag{1.2}$$

where we use the identifying restrictions $\lambda_0^A = \lambda_0^B = 0$. This independence model for two registers is denoted by [A][B].

There are two ways to derive the maximum likelihood estimate of the missed part of the population, $\hat{m}_{00}$: First, by using the estimate for the intercept such that $\hat{m}_{00} = \exp(\hat{\lambda})$ and second, by using the property that the odds ratio under independence is 1, i.e.,

$$\frac{m_{00}m_{11}}{m_{10}m_{01}} = 1, \tag{1.3}$$

so that,

$$\hat{m}_{00} = \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10}n_{01}}{n_{11}}. \tag{1.4}$$

1.2.2.1 Fully observed covariates

Covariates can be introduced to capture-recapture methodology with the aim to replace the strict independence assumption (Bishop et al., 1975). Covariates can also reduce the heterogeneity resulting from individual differences on that covariate (Alho, 1990). If covariates are available, the generally non-feasible independence assumption can be replaced by the less strict conditional independence assumption, where independence is assumed conditional on covariates (Bishop et al., 1975; Wolter, 1986a; Van der Heijden et al., 2012). This assumption is less stringent because it can take into account inclusion probabilities that are heterogeneous over the levels of the included covariate. Another advantage of using covariates is that it allows us to investigate the characteristics of the missing portion of the population.

Suppose we have observed covariate X, where the levels of X are indexed by x ($x = 1, 2$). Let m_{ijx} denote the expected values for A, B and X. Under independence conditional on X, there are two zero counts for cases not found in either register, namely a count for $x = 1$ and a count for $x = 2$.

Suppose that covariate X is the dichotomous covariate gender, with $x = 1$ is males and $x = 2$ is females. Let $n_x = n_{10x} + n_{01x} + n_{11x}$ and $\hat{N}_x = n_x + \hat{m}_{00x}$, where $\hat{m}_{00x}$ is defined as:

$$\hat{m}_{00x} = \frac{n_{10x}n_{01x}}{n_{11x}}. \tag{1.5}$$

Then the population size estimate assuming independence between A and B conditional on X, for $x = 1, 2$, is

$$\hat{N} = \sum_{x=1}^{2} n_x + \sum_{x=1}^{2} \hat{m}_{00x}. \tag{1.6}$$

The loglinear model for independence for two registers and covariate X is:

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX}, \tag{1.7}$$

where we use identifying restrictions $\lambda_0^A = \lambda_0^B = \lambda_2^X = \lambda_{01}^{AX} = \lambda_{02}^{AX} = \lambda_{12}^{AX} = \lambda_{01}^{BX} = \lambda_{02}^{BX} = \lambda_{12}^{BX} = 0$. When assuming independence between A and B conditional on X, $\lambda_{ij}^{AB} = \lambda_{ijx}^{ABX} = 0$. We denote this model as $[AX][BX]$.

1.2.2.2 Partially observed covariates

Sometimes data have partially observed covariates. Partially observed covariates occur when one register has a covariate that is not observed in the other register. Partially observed covariates can be approached as a missing data problem (Zwane and Van der Heijden, 2007; Van der Heijden et al., 2012). If we assume a Missing At Random (MAR) mechanism for the data, then we can use the Expectation-Maximization (EM) algorithm to estimate the missing values of the partially observed covariate of register A (and B) for the individuals not present in A (and B). MAR assumes that missingness depends only on the observed variables, and not on components that are missing (Little and Rubin, 2002, p. 12). When the MAR assumption has been satisfied the EM algorithm will give unbiased estimates.

Suppose register A has the covariate X_1, indexed by x ($x = 0, 1$), where the values for X_1 are missing for $A = 0$ because X_1 is not in register B. Assume that register B has the covariate X_2, indexed by y ($y = 0, 1$), where the values for X_2 are missing for $B = 0$ because X_2 is not in register A. The loglinear conditional independence model for two registers, with two partially observed covariates X_1 and X_2, is denoted as

$$\log m_{ijxy} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^{X_1} + \lambda_y^{X_2} + \lambda_{iy}^{AX_2} + \lambda_{jx}^{BX_1} + \lambda_{xy}^{X_1X_2}, \tag{1.8}$$

where parameters $\lambda_{ij}^{AB} = \lambda_{ix}^{AX_1} = \lambda_{jy}^{BX_2} = \lambda_{ijy}^{ABX_1} = \lambda_{ijx}^{ABX_2} = \lambda_{ijxy}^{ABX_1X_2} = 0$ when at least one of the subscripts is zero. The conditional independence model is denoted by $[AX_2][BX_1][X_1X_2]$. Inclusion of the parameter $\lambda_{ij}^{AX_2}$ instead of the parameter $\lambda_{ix}^{AX_1}$ may seem counterintuitive but an interaction for A and X_1 cannot be identified as the levels of X_1 do not vary over individuals for which $A = 0$, and similarly for B and X_2 (Zwane and Van der Heijden, 2007).

Table 2.6 illustrates that two registers, each with one dichotomous covariate, leads to 16 cells. However, because our covariates are only partially observed, for $B = 0$ columns $X_2 = 1$ and $X_2 = 0$ are collapsed, just as for $A = 0$ rows $X_1 = 1$ and $X_1 = 0$ are collapsed. In other words, we do not observe counts for m_{0111} and m_{0101} but only one count for the sum $m_{0111} + m_{0101}$, and similarly for $m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$ and $m_{1001} + m_{1000}$. Note that we have no observed values for m_{0011}, m_{0001}, m_{0010} and m_{0000}, as these refer to individuals who are in neither of the registers. Thus model $[AX_2][BX_1][X_1X_2]$ is saturated with eight observed values and eight parameters to be estimated.

TABLE 1.2
Expected values for two registers and two partially observed covariates.

		B = 1		B = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
A = 1	$X_1 = 1$	m_{1111}	m_{1110}	m_{1011}	m_{1010}
	$X_1 = 0$	m_{1101}	m_{1100}	m_{1001}	m_{1000}
A = 0	$X_1 = 1$	m_{0111}	m_{0110}	m_{0011}	m_{0010}
	$X_1 = 0$	m_{0101}	m_{0100}	m_{0001}	m_{0000}

1.2.3 Three registers

Population size studies can make use of multiple registers. The two register case can be easily extended to a three register case, which in turn can be extended to a d register case, with $d > 3$. Assume there are d samples, lists or registers. Assume also that within these sources, units are uniquely labeled or identifiable so that it can be determined in how many of the d sources a unit is present or absent. After correctly identifying these units as being present or absent in the d samples, the units can be identified as counts in a 2^d cross-classification such that there is a zero count for the units absent in all d samples (Bishop et al., 1975, P. 231).

Assume we have three registers, register 1, register 2 and register 3. Let variables A, B and C respectively denote inclusion in registers 1, 2 and 3. Let the levels of A be indexed by i ($i = 0$, 1) where $i = 0$ stands for ”not included in register 1”, and $i = 1$, stands for ”included in register 1”. Similarly, let the levels of B be indexed by j ($j = 0, 1$), and let the levels of C be indexed by k ($k = 0, 1$). Table 3.5 shows the expected values denoted by m_{ijk}. Observed values are denoted by n_{ijk}

with $n_{000} = 0$.

For three variables the saturated loglinear model is denoted by:

TABLE 1.3
The table of expected counts for thee registers

		C	
A	B	1	0
1	1	m_{111}	m_{110}
	0	m_{101}	m_{100}
0	1	m_{011}	m_{010}
	0	m_{001}	m_{000}

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}, \qquad (1.9)$$

where we use the identifying restrictions $\lambda_0^A = \lambda_0^B = \lambda_0^C = \lambda_{10}^{AB} = \lambda_{01}^{AB} = \lambda_{00}^{AB} = \lambda_{10}^{AC} = \lambda_{01}^{AC} = \lambda_{00}^{AC} = \lambda_{10}^{BC} = \lambda_{01}^{BC} = \lambda_{00}^{BC} = 0$. The model assumption is that the three factor interaction parameter $\lambda_{ijk}^{ABC} = 0$. Model $[AB][BC][AC]$ is the saturated model, as the number of observed counts equals the number of parameters to be estimated. This model assumes that the odds ratio between A and B is the same for $k = 0$ and $k = 1$, just as the odds ratio between A and C will be the same for $j = 0$ and $j = 1$ and the odds ratio between B and C to be the same for $i = 0$ and $i = 1$. Under this property

$$\frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}}. \qquad (1.10)$$

Then $\hat{m}_{000} = \exp\hat{\lambda}$ or, for the saturated model an estimate can easily be derived from (3.9), namely

$$\frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}} = \hat{m}_{000}. \qquad (1.11)$$

1.2.4 Assumptions

Capture-recapture methodology relies on at least five assumptions, however from the data it is not possible to verify whether they are met. Therefore it is important to know the effect of violated assumptions on the population size estimation. The five assumptions are:

1. Independence between the registers: for the two register case, the registers are assumed to be independent in the sense that the inclusion probability of register 1 is independent of the

inclusion probability of register 2. For three registers, this assumption is relaxed and it is only assumed that the three factor interaction is zero, such that dependence between pairs of registers may occur.

2. The registers are perfectly linked: when one unit is captured in two or more registers, perfect linkage assumes that we correctly identify all of these units as recaptures. Perfect linkage also means not linking units in two or more registers that do not belong to the same person, and thus should not have been linked.
3. The population is closed: for registers with continuous recording such as a Population Register the population is closed when one point in time is chosen. For incidence registers it is wise to take a small sampling period to limit a possible violation of the closed population assumption.
4. All individuals in the registers belong to the population, such that there are no erroneous captures, and
5. Assumptions related to homogeneity of inclusion probabilities (Van der Heijden et al., 2012).

Independence is the most researched assumption of capture-recapture methodology. The assumption of independence between two registers is very restrictive and can easily be violated. Under dependence between registers the inclusion probability of one register is related to the inclusion probability of the other register. Under positive dependence, an individual that is registered in register 1 has a higher probability of also being registered in register 2, resulting in an underestimation of the population size estimate. Similarly, under negative dependence, an individual in register 1 has a lower probability of also being registered in register 2, resulting in an overestimation of the population size estimate (Hook and Regal, 1993).

Capture-recapture methodology for human populations often makes use of existing data sources or survey samples that already exist (i.e. they are not created or collected with the aim to apply capture-recapture methodology). In using such sources there needs to be identification on every case in the data sources to achieve linkage. However, perfect linkage may be a difficult process. Registers may contain errors, or an already existing source is used that was not created for linkage, such that there may be little or no variables available to accurately identify the cases.

When the assumption of no erroneous captures is met, all individuals in the sample belong to the population. When there are cases in the data sources that do not belong to the population, capture-recapture analysis will estimate additional individuals that do not belong to the population. An older overview of different models to handle coverage errors in Censuses can be found in Wolter (1986a), more recent research on handling erroneous captures can be found in Zhang (2015).

When a population is open, individuals migrate in and out of the population during the sampling period. This migration during the sampling period may bias the estimator, given the data may not accurately describe the population. One way to keep the population "as closed as possible" is in using data sources that have units registered over a period of time. One time point can then be chosen to assess from the registers the number of individuals that are registered on that day. For data sources that are incident based, in most cases a period of time has to be chosen to get a decent sample of the population. This could be a week, a month, six months, or a year, depending on known migration flows and the amount of uncertainty the researcher is willing to take. The longer the period chosen, the larger the size of the violation of closure is likely to be. However, often, when the sampling period is taken too small, there are not enough units in the sampling period to make a precise estimation of the missed part of the population. Capture-recapture analysis has been applied to open populations and has been adapted to accommodate open populations. An interesting overview can be found in Chapter 17 of Lawless (2014), where also applications to data involving humans are discussed.

Registers may have heterogeneous inclusion probabilities, for example when the probability to include men is higher than the probability to include women. If there is one source of heterogeneity, the population size estimate is unbiased when at least for one of the two registers the inclusion probabilities are homogeneous (Chao et al., 2001; Zwane and Van der Heijden, 2007; Van der Heijden et al., 2012). If there is a source of heterogeneity in each of two registers, the estimates are unbiased if the inclusion probabilities of the two sources of heterogeneity are statistically independent (Seber, 1982; Van der Heijden et al., 2012).

1.3 Practical background information for this thesis

1.3.1 Census

Every 10 years, every European Union country provides a Census count to Eurostat. Censuses were traditionally based on full field enumeration, where information was collected using paper census forms (United Nations Economic Commission for Europe , UNECE). Enumeration officials would go from door to door to enumerate each individual or household to interview via a paper Census form. However, full field enumeration is expensive and time consuming. Some countries still using the traditional Census forms, such as Estonia, have in part also used online forms for the 2011 Census round. Other countries such as the US, Australia and New Zealand aim to incorporate online Census forms into their next Census round.

The nordic countries were one of the first to start using administrative data for the Census, where Denmark was the first to use a fully register-based Census in 1981 (United Nations Economic Commission for Europe , UNECE). For the 2011 Census round, 9 of the UNECE region countries are using a full register-based Census, of which the Netherlands is one. The Netherlands have not used a full enumeration Census since 1971 and now relies on administrative data and one survey for its register based Census (Statistics Netherlands, 2015).

For the Dutch Census enumeration, the Population Register (PR) is used, which for the 2011 Census round was still the Gemeentelijke BasisAdministratie (GBA, and is currently replaced by a BRP, Basis Registratie Personen). To assess the quality of the register an estimate of the undercoverage is needed. For that purpose we can use capture-recapture methodology. The method includes linking the individuals in the registers and subsequently estimating the number of individuals missed by both registers.

There is literature going back to the 1940s under the title dual system estimation or dual record systems using a two sample capture-recapture method for the Census. In countries with a traditional Census, a Post-Enumeration Survey (PES) could be organised to collect recaptured data, as was the case for instance in the United Kingdom (Brown et al., 1999, 2006; ONS, 2012b), and in the US (Wolter, 1986a,b; Bell, 1993). An interesting overview of capture-recapture methodology on the US Census and adjustment for Census undercount can be found in Fienberg (1992). It is of interest to note that there is a higher probability that independence is met when linking a PES to the Census, rather than using two registers. However, this is only the case when the Census individuals or addresses are not used as a sampling frame.

Under Dutch regulations, every individual residing in the Netherlands for longer than four months, or is planning to do so, has to register in the PR. Thus, the PR contains demographic information on the de jure population, which differs from the 'de facto' population that is defined as the Dutch population which encompasses, but is not restricted to the PR. This incompleteness of the PR has more than one reason. First, within the European Union there is free movement and employment for individuals with a European Union nationality, such that individuals that have not registered themselves in the PR are legally residing in the Netherlands but will be missing from the PR. Second, the PR is incomplete due to undocumented immigrants, coming from outside the European Union without a working or residence permit, in most cases, formerly asylum seekers.

Assessing the undercoverage of a register asks for a definition of the Dutch population. According to the United Nations Statistics Division (2008, p. 102) Principles and Recommendations for Population and Housing Censuses Rev.2, we can define the population of a European country along the terms of usual residence:

"1.461. In general, "usual residence" is defined for census purposes as the place at which the person lives at the times of the census, and has been there for some time or intends to stay there for some time",

where usual residence is defined as:

"1.463. It is recommended that countries apply a threshold of 12 months when considering place of usual residence according to one of the following two criteria: (a) The place at which the person has lived continuously for most of the last 12 months (that is, for at least six months and one day), not including temporary absences for holidays or work assignments, or intends to live for at least six months; (b) The place at which the person has lived continuously for at least the last 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months."

The European Union generated similar definitions, which can be found in Regulation (EU) No 1260/2013 of the European Parliament. In the EU, usual residence is defined similar to 1.463.b of the United Nations Statistics Division (2008, p. 102, 103), as:

"The place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits

to friends and relatives, business, medical treatment or religious pilgrimage".

An individual is considered a usual resident when they have lived in the Netherlands for a continuous period of 12 months before the reference time, or if they arrived in the 12 months before the reference time and intend to stay for at least a year. When these circumstances can not be established, "usual residence" means the place of registered residence. The registers used in this thesis may register a form of residence duration based on registration date, but not on intent to stay. Intent to stay is not commonly documented in registers, unless specifically asked. Hence, in this thesis only a length of stay will be used to assess usual residence instead of an intent to stay.

1.3.2 Data used

For the two register case, a Crime Suspects Registers (CSR) has been linked to the Dutch Population Register (PR). The PR, registers date of registration and date of death or immigration, such that it contains a period in which a person is registered as residing in the Netherlands. The CSR is an incidence register on suspects of known crimes. However a case in the CSR is a report on a crime filed by the police, rather than registering persons. Thus on a given day only a few reports may have been filed and for the CSR we can not take one date to get the registered population. Rather, a period of time has to be used, in this case of six months. The covariates used for the two register case have been age, marital status or police region. The data used is either from 2007 or from 2009. In chapter 3 of this thesis the data used is from the PR and CSR of the year 2010. The data are from different years because over time, data became available from more recent years and the more recent data was used.

For the remaining chapters in this thesis three registers have been used. These registers are the PR and CSR with the addition of an Employment Register (ER). The ER is a register that does not document individuals but jobs. For the purpose of our analyses the job-register of 2010 has been transformed into a register on individuals. Jobs were attributed to the individuals holding those jobs, so that the ER can be transformed to a database on individuals. In the three register case we used the covariates usual residence, sex, age and nationality.

1.3.3 Linking the data

Two types of linkage have been used in this thesis to link the three registers. Deterministic linkage considers two individuals in a pair of registers a link when they agree on a linkage key. An example of linkage keys is a Personal Identification Number (PIN), or a combination of variables such as address, date of birth, etc.

At statistics Netherlands all registers are linked deterministically to the PR via a PIN and a linkage key. The PR is the backbone of Statistics Netherlands, such that all registers and surveys are linked to the PR. Linking other registers, of which also the ER and the CSR, to the PR via a Dutch PIN, enables about 96 to 98 percent of the cases to be linked to the PR. When a case has no PIN, the registers are linked on postal code, house number, date of birth and sex. Then 93 to 95 percent of individuals can be linked to the PR (Arts et al., 2000).

Small errors can be accounted for in deterministic linkage, such as spelling errors or errors that occurred when the data was entered into the register. However for those individuals in the ER and CSR that were unable to be linked deterministically, either to one another or to the PR, the information on the linkage key contained bigger errors that made deterministic linkage difficult. To further improve upon deterministic linkage the individuals in the registers are also linked pairwise via probabilistic linkage.

Fellegi and Sunter (1969) mathematically formalized probabilistic linkage, where pairs of records are classified as either a link, a non-link or possible links (Herzog et al., 2007). For probabilistic linkage, probable links are created for all cases in a pair of registers. The probability of a link is created for each possible pair on each element of the linkage key, for example a string of variables, such that each pair results in a numerical value of their similarity. This numerical value of their similarity is called a weight (Herzog et al., 2007). A cut-off can be determined above which weights the researcher considers each pair a link, a non-link or a possible link. Probabilistic linkage enabled us to link the ER and the CSR, and also remaining ER and CSR cases to the PR. However, during linkage it was found that 37 % of the units that were registered in the CSR had missing linkage key information and therefore were unable to link to the PR and ER.

1.3.4 Usual Residence

One very important covariate in this thesis is usual residence. The United Nations Statistics Division (2008) defines a country's population along the terms of usual residence, which is used in this thesis to estimate the undercoverage of the Dutch PR. I only look at undercoverage of the PR in this thesis, because over coverage is less than a problem. Bakker (2009) estimated an over coverage of 31 thousand individuals.

Unfortunately, neither register has a measure of residence duration. For two of the three registers residence duration can be deduced. The PR has a registration date, the date at which either a person was born into the Netherlands or immigrated into the Netherlands and was officially registered as a Dutch resident. Then the Census date minus the registration date can be used as a residence duration. This residence duration is in days and can be transformed into a categorical variable for either residing in the Netherlands for shorter or longer than a year.

For the ER, the information available from which to deduce residence duration are the job lengths. The ER registers the start and end date of jobs held by all individuals registered in the ER. When a person had only one job, the start and end date of this job can be used as a residence duration. When a person has held two or more consecutive or overlapping jobs, these can be merged to one residence duration. There were however individuals in the ER that had multiple jobs that were not consecutive, where in between two jobs for a specific period of time, no job was held. It is reasonable to assume that a specific period of time between jobs will mean that the person still resided in the Netherlands and was in between jobs at that time.

Because we do not know which period of time between jobs is reasonable to assume that the person was still residing in the Netherlands, different scenario's were conducted to investigate the effect that different periods of time allowed between jobs had on the number of usual residents. From this analysis it was found that 31 days of unemployment between two jobs was a reasonable time period to still consider two jobs as one consecutive residence duration. When an individual leaves the country for a maximum of a month between jobs it is more likely a holiday, and even if the individual leaves to another country they are still in the Netherlands for the remaining 11 months. Additionally, a month of unemployment can financially be bridged. From the analysis it was also found that allowing a larger period of unemployment increased the number of usual residents rapidly, which was deemed unrealistic.

The CSR has no information on residence duration. Therefore, for those individuals that did not link to the PR and the ER there is missing information. In this PhD thesis, two ways of handling the missing usual residents variable has been used. These two methods are the Expectation Maximization (EM) algorithm and Predictive Mean Matching (PMM) multiple imputation.

1.4 Contribution of this thesis

The current thesis consists of four chapters all concerning capture-recapture estimation on Dutch data. Sensitivity analyses and a simulation study have been set up to assess the effect that violated assumptions have on the population size estimation. Most interesting is the finding that the effect of a violation of any assumption has on the bias of the population size estimator is due to the implied coverage of the registers. Additionally it was investigated which method best handles missing data in capture-recapture methodology. The information that has been gained from this research has been used in the second part of this thesis, where we estimate the undercoverage of the Dutch PR.

1.4.1 Implied coverage

One overarching finding in this thesis is the effect of implied coverage on the population size estimator. This finding will be discussed first. An important property of implied coverage is that it can be estimated from the data, whereas it can not be ascertained from the data whether assumptions are met. Thus, implied coverage gives researchers an indication whether their population size estimate will be robust under violations of the assumptions of capture-recapture.

Under register 1 and 2, the maximum likelihood estimate of the missed portion of the population can be estimated via equation (3.2). Under (3.2) we can estimate conditional probabilities:

$$\hat{p}(0|1) = \frac{n_{01}}{n_{+1}} \text{ and, } \hat{p}(1|1) = \frac{n_{11}}{n_{+1}}, \tag{1.12}$$

where $\hat{p}(0|1)$ is the estimated probability of new cases provided by register 2. Similarly, $\hat{p}(1|1)$ is the estimated probability that cases in register 2 are already known in register 1. Given these probabilities we can rewrite equation (3.2) as

$$\hat{m}_{00} = \frac{n_{10} * \hat{p}(0|1)}{\hat{p}(1|1)} = \frac{n_{10} * \hat{p}(0|1)}{1 - \hat{p}(0|1)}. \tag{1.13}$$

It can be seen from equation (3.6) that the estimated number of individuals missed by the two registers is a function of the estimated probability of new cases added by register 2. When the estimated probability of new cases $\hat{p}(0|1)$ is relatively small, we say that the coverage of register 1 implied by register 2 is high. This is further referred to as "implied coverage". However, when the estimated probability of new cases $\hat{p}(0|1)$ is relatively large, then the coverage of the population by register 1, implied by register 2 is low, such that register 2 captures a relatively larger number of unique cases compared to register 1.

When the implied coverage is high, violations of the assumptions will usually make minor modifications to $\hat{p}(0|1)$ and therefore will have a small effect on the population size estimate. When implied coverage is low, the population size estimation is less robust to possible violations of the assumptions. This follows from (3.2).

Choosing which register is register 1 and which is register 2 depends on the size of the registers used. The register containing most cases will be denoted by register 1 and the smaller register will be denoted by register 2.

Implied coverage for three registers

Implied coverage can be extended to the three register case. Assume that register 1 covers most of the population, then register2 and then register 3. For our purposes it suffices here to discuss coverage of registers 1 and 2 implied by the third register. For this purpose Equation (5.2) is complicated, as n_{111}, the number of cases seen in all three registers, is in the numerator. We focus on the observed counts n_{101}, n_{011} and n_{001}, i.e. the number of individuals seen only in register 3, i.e. n_{001}, compared to the individuals seen in register 3 and only in register 1, i.e. n_{101}, and compared to the individuals only seen in register 3 and only in register 2, i.e. n_{011}. We focus first on n_{001} and n_{101}. Notice that these counts refer to individuals missed by register 2. Therefore we will speak of coverage conditional on being missed by register 2. Now, similar to the discussion of implied coverage in a two-way table above, if n_{001} is large in comparison to n101, the conditional coverage of register 1 implied by register 3 is low, where conditional refers to being missed by register 2. Thus the estimator in equation (5.2) becomes unstable when the number n_{001} becomes unstable. Similarly, for n_{001} and n_{011}, if n_{001} is large in comparison to n_{011}, the conditional coverage of register 2 is low and the estimator in equation (5.1) becomes unstable when the number n_{001} becomes unstable (where conditional refers to being missed by register 1. Anyhow, for all practical purposes it will be clear that when n_{001} is large, the estimator defined in (5.2) will become unstable.

1.4.2 Sensitivity of population size estimation for violating parametric assumptions in loglinear models

In chapter 2, the impact that mild or severe violations of (conditional) independence have on the population size estimate is investigated. Out of the five assumptions discussed above, dependence in capture-recapture analysis has been researched the most. Research to the impact of the violation of independence usually involves simulation studies, an already known population size estimate or multiple sources (Wolter, 1986a; Hook and Regal, 1992; Bell, 1993; Hook and Regal, 1997; Cormack et al., 2000; Hook and Regal, 2000; Baffour et al., 2013).

In this thesis an extension has been made from the work of (Brown et al., 2006), who conducted sensitivity analyses to study the effect of dependence on real data. The odds ratio under independence is $m_{00}m_{11}/m_{10}m_{01} = 1$. When independence is not met, $m_{00}m_{11}/m_{10}m_{01} \neq 1$, and we are operating under dependence. We can denote dependence using θ such that $m_{00}m_{11}/m_{10}m_{01} = \theta \neq 1$ and dependence can be simulated. The maximum likelihood under dependence of size θ can be estimated via

$$\hat{m}_{00(\theta)} = \theta\frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \theta\frac{n_{10}n_{01}}{n_{11}} = \theta\hat{m}_{00}. \tag{1.14}$$

The results from this chapter show that for two different nationality groups, the population size estimate under dependence could be fairly robust as well as not robust at all. For the Afghan, Iraqi and Iranian people the population size estimate did not change much when dependence was introduced; it also remained fairly robust whether or not we assumed conditional independence on fully observed covariates. However for the Polish people, the population size estimate changed dramatically when dependence was introduced, whether or not conditional independence was assumed on fully observed covariates.

The difference between these nationality groups is that for the Afghan, Iraqi and Iranian group, implied coverage is high, such that the population size estimator is robust against dependence, whereas for the Polish group, the implied coverage is rather low and the population size estimator is not robust against dependence.

This chapter also investigated partially observed covariates. Researchers are often faced with covariates that are only found in one register. Such a covariate may be of high importance to the research question. Usually these partially observed covariates are ignored. Research had shown how partially observed covariates can still be used in estimation (Zwane and Van der Heijden, 2007; Van der Heijden et al., 2012), and in this chapter that work has been further elaborated. Dependence between the register and the partially observed covariate did not impact the population size estimation much, concluding that the population size estimator is robust against dependence between registers and partially observed covariates.

1.4.3 The effects of imperfect linkage and erroneous captures on the population size estimator

In chapter 3 the effect of violated assumptions on the population size estimate in capture-recapture analysis is further investigated. In the previous section the independence assumption was investigated; in this chapter the effect of violation of the assumption of perfect linkage and no erroneous captures will be investigated. The interest lies especially in the effect that violations of these assumptions have on the population size estimate and whether that effect is the result of implied coverage. As far as we know, research into the effect of implied coverage on linkage error and erroneous captures has not been done before.

First linkage error is discussed. Unfortunately, we can not verify from the data to what extent the assumption of perfect linkage has been violated. We can however investigate the effect of linkage error on the population size estimate in a sensitivity analysis. Assume we have linkage errors of size b, where b is the number of false positive links minus the number of false negative links. Then b is negative when the number of false negative links outbalance the number of false positive links, and b is positive when the number of false positive links outbalance the number of false negative links. Under perfect linkage $b = 0$ for observed values n_{ij}. Under linkage error $b \neq 0$, observed values are denoted $\tilde{n}_{ij}$. Then $\tilde{n}_{11} = n_{11} + b$, $\tilde{n}_{10} = n_{10} - b$ and $\tilde{n}_{01} = n_{01} - b$.

To choose values of b we define linkage error rate β for n_{01}. Linkage error rate β has been chosen for n_{01} because n_{01} is the number of added cases of register 2 relative to register 1, such that β is specified in relation to the implied coverage of register 1 given register 2. Then,

$$\beta = \frac{\tilde{n}_{01}}{n_{01}} \tag{1.15}$$

where $\beta = 1$ denotes perfect linkage. Linkage error rate β enables us to simulate linkage error, where $\tilde{n}_{01} = n_{01} * \beta$. In creating such a linkage error rate β we can denote linkage error in percentages, and by defining linkage error in percentages we can better compare the effect of β on the population size estimate between the two nationality groups. Then, when $\tilde{m}_{ij}$ is the expected values related to $\tilde{n}_{ij}$,

$$\frac{\hat{\tilde{m}}_{10}\hat{\tilde{m}}_{01}}{\hat{\tilde{m}}_{11}} = \frac{\tilde{n}_{10}\tilde{n}_{01}}{\tilde{n}_{11}} = \frac{(n_{10} - b)(n_{01} - b)}{n_{11} + b} = \hat{m}_{00(\beta)}, \tag{1.16}$$

where $\hat{m}_{00(\beta)}$ is the size of the individuals missed by the two registers. The population size estimate under linkage error is $\hat{N}_{\beta} = \hat{m}_{00(\beta)} + (n_{11} + b) + (n_{10} - b) + (n_{01} - b)$.

Second, we investigate the effect erroneous captures have on the population size estimator. It is not always possible to verify from the data whether cases are erroneous captures, but we can simulate erroneous captures to assess their effect on population size estimation. For n_{ij} we assume no erroneous captures. If erroneous captures are introduced, observed values under linkage error are $\bar{n}_{ij}$. We can define an erroneous capture rate γ for n_{01}, where $\gamma = \bar{n}_{01}/n_{01}$, such that $\bar{n}_{01} = n_{01} * \gamma$. Erroneous captures are units in the data that should not have been observed, and therefore $\bar{n}_{01}$ will always be smaller than n_{01}, and $0 \leq \gamma \leq 1$. Erroneous capture rate γ has been defined on n_{01} because that is the number of added cases by register 2, relative to register 1, and is related to the implied coverage. We find

$$\hat{m}_{00(\gamma)} = \frac{n_{10}(n_{01*}\gamma)}{n_{11}} = \frac{n_{10}(\bar{n}_{01*}\gamma)}{n_{11}} = \gamma\hat{m}_{00}, \tag{1.17}$$

where $\hat{m}_{00}$ is the estimate when there are no erroneous captures defined in (3.2).

Again two data sets are compared. The data of Afghan, Iraqi and Iranian individuals residing in the Netherlands in 2010 are compared to data of Polish individuals residing in the Netherlands in 2010. We compare the same data as in the previous section, with the exception that it is from another year. We have defined a general and flexible linkage error rate and erroneous capture rate that can be specified to one linkage error rate and erroneous capture rate that can be used for a sensitivity analysis.

As in Chapter 2, for the Afghan Iraqi and Iranian individuals, the population size estimator was relatively robust, whereas for the Polish individuals the population size estimator was not robust at all to violations of assumptions. In this chapter it was found that implied coverage of the registers is an important aspect in why some nationality groups have a robust population size estimator and why some nationality groups do not have a robust population size estimator. Implied coverage, as described above, was found to be the determinant in whether violations of perfect linkage and no erroneous captures have a large effect on the population size estimator.

Additionally, an extension has been made to the three register case. To assess the effect of linkage error on the three register capture-recapture estimation a simulation study has been conducted using the same properties in the simulated registers as the actual data. The simulation study resulted in an indication where the most linkage errors could be found, such that for the actual data this could be addressed. The sensitivity analyses to the effect of linkage error and erroneous capture on the population size estimator showed the same result as for the two register case. The population size estimator using three registers is robust to possible violations of the assumptions for the Afghan, Iraqi and Iranian individuals, but not for the Polish individuals.

1.4.4 Different methods to complete datasets used for capture-recapture estimation: estimating the number of usual residents in the Netherlands

Chapter 4 investigates which method is best to handle missing data introduced by partially observed covariates. As discussed before, for the Crime Suspects Register (CSR) there was no measure on usual residence. However, to estimate the undercoverage of the PR we need an estimate of the usual residents. Therefore, usual residence has to be completed for the CSR.

There are different methods available to handle missing data. In this chapter we use the Expectation Maximization (EM) algorithm and Predictive Mean Matching (PMM). The EM algorithm is often used in categorical data analysis, also in previous research concerning partially observed covariates (Zwane and Van der Heijden, 2007; Van der Heijden et al., 2012). PMM has the advantage of flexibility in the choice for a specific part of the observed data that will be used for the imputation of the missing data. Four scenarios have been identified where the missing data are completed via either the EM algorithm or PMM imputation, resulting in different population size estimates for usual residence. The different scenarios lead to different population size estimates; even small changes in the completed data lead to differences in plausibility of the estimates. In this study PMM imputation performs better in terms of flexibility and plausibility of the estimates.

1.4.5 Sensitivity of the population size estimates for Census undercoverage

For countries using a register based Census, an estimate of the undercoverage of the registers is important information, and is necessary when estimating the number of usual residents for the Census. However, the

PR only contains information on the registered population. Therefore, the PR alone is not sufficient to estimate the number of usual residents. This chapter documents the case of the Netherlands in estimating the usual residents via capture-recapture methodology for the undercoverage of the PR. This research builds upon the information gained from the earlier chapters in this thesis, which is used to exemplify how researchers can deal with practical challenges in the use of administrative data for population size estimation.

A couple of challenges arose during the research process. It appeared that neither register had a measure of usual residence, and additional measures had to be taken to complete the incomplete usual residence variable. Also, for each capture-recapture assumption extra measures had been taken to make sure they were met as best as possible. However, it was found during the linkage process that 37% of the individuals in the CSR that did not link to the PR and the ER had no, or incomplete, linkage key information. It was uncertain whether these individuals belonged to the population, nor whether they still had to be linked to the PR and/or ER. Therefore, different scenario's have been set up where different percentages of linkage error and erroneous captures were simulated.

From these scenarios a range of the undercoverage of the PR is given. The undercoverage of the PR ranges from 88 to 185 thousand usual residents aged 15 to 65, which means that we have an undercoverage of the PR of only .5 to 1.1 percent. Given that this undercoverage does not include children up to the age of 15, and elderly over 65, the real undercoverage will lie somewhat higher This overlapped with a range based on previous research. Based on this research, the number of usual residents is expected to lie between 175 and 225 thousand individuals. Due to this overlap we expect that the estimate of the undercoverage will most likely lie in the upper end of the range of 88 to 185 thousand usual residents.

1.4.6 Overall conclusion

This thesis has provided new knowledge on capture-recapture methodology to answer the main questions posed in section 1. Question 1 asked "what is the effect of violated assumptions and missing data on the robustness of the estimation of population size estimation via capture-recapture methodology?". Capture-recapture methodology relies heavily on a couple of assumptions. For three of these assumptions sensitivity analyses have been conducted, i.e. the independence assumption, perfect linkage and no erroneous captures. To compare, we used data on two nationality groups: a group of individuals with an Afghan, Iraqi and Iranian nationality and a group of Polish individuals. It was found that implied coverage impacts the effect violated assumptions have on the population size estimation. The difference in implied coverage of the two nationality groups have the effect that for the individuals with an Afghan, Iraqi and Iranian individuals the estimator is relatively robust to violating the assumptions tested in this chapter. However for the Polish individuals a violation results in seriously biased results. This result is important, because implied coverage can be estimated, whereas the extent that assumptions may have been violated can not.

The second question asked "how can the information gained in the first question be used to achieve a trustworthy estimate of the under coverage of usual residents in the Population Register at Statistics Netherlands"? A few examples are given in this thesis of possible practical challenges and how it has been chosen to handle these. The most common challenge in the data is missing information. In chapter 2 of this thesis it is shown that using the EM algorithm, partially observed covariates can still be used for capture-recapture estimation. We elaborate on Zwane and Van der Heijden (2007) and Van der Heijden et al. (2012) by investigating what would happen when there is dependence between a partially observed covariate and the register. It has been shown in this chapter that dependence between partially observed covariate and register has little effect on the resulting population size estimate. Thus the EM algorithm can be used without caution to estimate the missing values of a partially observed covariate when these are missing in one register.

In chapter 4 of this thesis missing data for covariates are further discussed. Different scenario's of missing data methods have been investigated to complete the missing covariate usual residence. For the individuals in the CSR that did not link to the PR or the ER we had no indication of their residence duration. Four scenarios were set up, of which 3 used the EM algorithm under different loglinear models and one scenario had Predictive Mean Matching (PMM) multiple imputation. A slight advantage was found for PMM imputation, because it has the flexibility of choosing which subgroup to use as donor information for the missing data.

Chapter 5 uses the information gained from the earlier chapters to estimate the undercoverage of the PR. The thesis concludes with two overarching findings. First, implied coverage of the data is important. The effect of whether a violated assumption has a large effect on the population size estimator, or not, is a direct result of the implied coverage. Therefore researchers are advised to assess the coverage of their registers, given that implied coverage will dictate the effect possible violations will have on the estimation. Second, the undercoverage of the PR for individuals aged 15 - 65, will most likely be in the upper end of the range of 88 to 185 thousand individuals, resulting in an undercoverage of only 0.5 to 1.1%.

1.4.7 Future research

In this thesis the effect of violating three specific assumptions on the population size estimator has been investigated. Two other assumptions have not been studied, i.e. homogeneity and the closed population assumption. It was found that implied coverage plays a role in the effect that violation of the assumption has on the population size estimator, and to complement the research from this thesis it would be interesting to investigate the effect of heterogeneity and an open population.

Erroneous captures have been discussed in this thesis, and especially the effect that erroneous captures can have on the population size estimator. When individuals do not belong to the population but are observed in the data, they will bias the population size estimator. Future research should focus on better identifying individuals that do not belong to the population such that they can be deleted before estimation. There are few advances in this field, for one, the Trimmed Dual System Estimator of Zhang and Dunne (2015) and Zhang (2015). They have developed a method to estimate the number of the erroneous captures in the data. Based on known information, the cases with the highest probability of being an erroneous capture are removed. They assume that when the population size estimate decreases, it means that the researcher correctly identified erroneous captures to be removed. According to them, when the population size estimate goes up or flatlines, this indicates the unit deleted belonged to the population and was not allowed to be removed. They find it is important that cases are not deleted randomly. This method has been applied to the Irish population data, and its performance on the Dutch population should also be studied.

In administrative data, there are often errors that make record linkage difficult. It was found that linkage errors can have a large biasing effect on the population size estimator, and for future research it is advised to improve further upon linkage to reduce bias in the population size estimates. An interesting advancement in record linkage research is the research by Consiglio and Tuoto (2015). They adjust the Petersen estimator by explicitly taking into account linkage errors. Their research overcomes the limitations in the work by Ding and Fienberg (1994), who proposed a method for unbiased estimates when there is linkage error. By defining the probabilities of being counted in both lists, Consiglio and Tuoto (2015) improved upon the Ding and Fienberg estimator. It would be interesting to implement their Modified Ding and Fienberg (MDF) estimator on data by Statistics Netherlands, and also to extend the MDF to the three register case.

Three registers have been used in this thesis, the PR, the ER and the CSR. The PR is the Dutch population register and already covers most of the population. The ER covers the working population and the CSR can in theory cover the whole of the Dutch population, given that every individual could be suspected of a crime. It may however be that other registers cover the population better, or may be able to cover a specific subpopulation especially missed by the other registers. It would therefore be interesting to conduct capture-recapture analysis on other datasources. It may be possible that data on healthcare or hospitals have better coverage of the population than one of the currently used registers. Every individual will need medical care and thus may end up in such a register, which may improve heterogeneity if these individuals are documented correctly. An interesting sample frame may be one with addresses instead of persons such as the PR. Using an address register, individuals may be found that are residing in the Netherlands but who are not registered in the PR.

In this thesis no coverage rates were available. These were however available in the research of Bryant and Graham (2015), where they used a coverage survey to estimate coverage rates per regions. These coverage rates were then added as an implicit weight to the model and improved the population size estimator. Their research highlight the importance of coverage surveys. For the Dutch case, it may be interesting to conduct a coverage survey or use information from migration flows to improve upon estimation.

Capture-recapture methodology is a useful method for Official Statistics and improvements of the methodology will greatly aid in estimating the size of a countries population needed for the Census. This thesis has investigated a couple of aspects of capture-recapture estimation and covered some practical problems that may be found in applying capture-recapture methodology to Official Statistics. Additionally, some suggestions have been made for future research to improve further on population size estimation.

2

Population size estimation after violating parametric assumptions

Susanna C Gerritse

Statistics Netherlands

Peter G.M. van der Heijden

Utrecht University/University of Southampton

Bart F.M. Bakker

Statistics Netherlands/VU University

CONTENTS

This chapter has been published as: S. C. Gerritse, P. G. M. Van der Heijden, and B. F. M. Bakker. Sensitivity of population size estimation for violating parametric assumptions in loglinear models. Journal of Official Statistics, 31:357 - 379. 2015. We thank Aat Liefbroer, Ray Chambers and the CBS Advisory Committee on Census

2.1 Introduction

For the Census of 2011 an increasing number of countries used administrative data to collect the necessary information. Under Census regulations a quality report is obligatory, and one of the aspects that needs to be addressed is the undercoverage of the Census data. This asks for an estimate of the size of the population. If one wants to estimate the size of a population, capture-recapture methods, making use of loglinear models, are commonly used (Fienberg, 1972; Bishop et al., 1975; Cormack, 1989; International Working Group for Disease Monitoring and Forecasting, 1995). These methods go by different names, such as mark-recapture, dual system methods or dual-record system methods. In this paper we use the label capture-recapture. In countries with a traditional Census a post-enumeration survey could be organised to collect recaptured data, as was the case for instance in the United Kingdom (Brown et al., 1999; ONS, 2012b), and in the US (Wolter, 1986a; Bell, 1993; Nirel and Glickman, 2009). In that case a survey with a relatively small sample size is linked to the Census data. In countries with a Census based on administrative data, the approach mostly used is to find two registers and treating these as the captured and recaptured data. The method includes linking the individuals in the registers and subsequently estimating the number of individuals missed by both registers.

However, the outcome of the capture-recapture method depends heavily on some assumptions underlying the data. In particular, if two sources are used, it is usually assumed that inclusion in the captured data is independent of inclusion in the recaptured data. A second assumption deals with homogeneity versus heterogeneity of inclusion probabilities. If there is one source of heterogeneity it is assumed that at least for one of the two sources the inclusion probabilities are homogeneous (Chao et al., 2001; Zwane and Van der Heijden, 2004). If there are two sources of heterogeneity (two covariates) the estimates are unbiased if the inclusion probabilities of the first source vary with one source of heterogeneity, and the inclusion probabilities of the second source vary with a second source of heterogeneity, but the two sources of heterogeneity are statistically independent (Seber, 1982, p. 86). The remaining two assumptions are that the population is closed and that the registers are perfectly linked.

Under Coverage for their valuable comments and an anonymous reviewer for coining the term implied coverage. Acknowledgement of author contributions: theoretical development by all authors. Document written by SG, edited by PvdH and BB. Analyses carried out by SG.

The assumption of independence between two registers is very strict and can easily be violated. Under dependence between registers the inclusion probability of one register is related to the inclusion probability of the other register. Then, under positive dependence individuals in the captured data have a higher probability of also being in the recaptured data, resulting in an underestimation of the population size estimate. Additionally, under negative dependence the opposite holds (Hook and Regal, 1995).

Independence is an unverifiable assumption, i.e., it cannot be verified from the data used for the estimation of the population size. The loglinear independence model for the linked captured and recaptured data has three parameters whereas there are only three counts. Because the observed counts are equal to fitted counts, the independence model is the saturated model (compare, Van der Heijden et al., 2012). Thus we cannot assess dependence from the saturated model. One way of reducing the impact of the strict independence assumption is to replace it with the lesser strict assumption of independence conditional on covariates. Adding covariates enables us to reduce heterogeneity introduced to the model due to the specific covariate, adjusting the population size estimate for the better. The situation of a saturated model also holds when covariates of individuals are taken into account and we operate under the loglinear conditional independence model. Yet we are interested in what the impact of mild or severe violations of (conditional) independence is on the population size estimate. It does not necessarily have to be the case that violation of the (conditional) independence assumption results in a substantive bias in the population size estimate. It is of importance to also assess what happens when the other assumptions are violated. However, looking at all assumptions at once is very complex. Thus in this paper we will focus on the violation of the independence assumption, assuming all other assumptions to be met.

We propose a general approach to sensitivity analyses under the loglinear model framework using a loglinear Poisson regression, a special case of the generalized linear model. Where in the saturated model specific interaction parameters are equal to zero, we impute fixed values departing from zero for these parameters, thus simulating dependence, and investigate the impact on the population size estimate. As the loglinear interaction parameters are closely related to (conditional) odds ratio, there is a clear interpretation for the values to which we fix the parameters.

Similar findings come from the research of Brown et al. (1999) where the Census was linked to a Post Enumeration Survey to assess under and overcoverage (compare also, Wolter, 1986a; Bell, 1993). Brown et al. (1999) used a fixed odds ratio of 0.1 and 10 to investigate the impact of simulated dependence on the population size estimate. They showed that fixed dependence can seriously bias the population size estimate under the independence assumption. Results like these are valuable since they give insight to the size of the impact of violated independence. However, research into the robustness of the population size estimator under violation of independence is non standard. As far as we know, other research to the impact of the violation of independence involves simulation studies, an already known population size estimate or uses multiple sources (Wolter, 1986a; Bell, 1993; Cormack et al., 2000; Hook and Regal, 1992, 1997, 2000; Brown et al., 2006; Baffour et al., 2013).

We extend the results of Brown et al. (1999) by, instead of using the standard loglinear model, working under a loglinear Poisson regression where we simulate a fixed dependence using offsets. In simulating dependence by adding a fixed offset value to the loglinear model we can compare the population size estimate under independence to the population size estimate under a 'true' dependence. Additionally we extend our two register independence model to the case with covariates observed in both registers (fully observed covariates) and covariates observed in only one register (partially observed covariates).

Partially observed covariates are usually ignored because including them would lead to missing values in the other register. However ignoring these covariates when they actually are related to the inclusion probability of the register results in a biased population size estimate (Zwane and Van der Heijden, 2007). In assuming MAR we can impute the missing values of the partially observed covariate in the other register and use this covariate to replace the strict independence assumption with independence conditional on covariates. For partially observed covariates the loglinear model is easily extendable so that we can also conduct sensitivity analyses in this context.

We proceed as follows. In section two we will discuss the loglinear model for a capture-recapture model with two registers without covariates. In section three we will discuss a two register capture-recapture model and conduct a sensitivity analysis on two registers with a conditional independence. In section four the independence assumption will be conditional on partially observed covariates, where a covariate has been observed in only one register. Here the sensitivity analysis is on the dependence of the partially observed covariate on the register, thus whether the covariate influences the inclusion probability of the register. Section five provides some extensions made to a model, namely for models for three registers, the multiplier method and confidence intervals.

We use two datasources to illustrate the robustness of capture-recapture methodology, that have been provided by Statistics Netherlands. We chose not to make a simulation study because researchers in the field of capture-recapture use real data and we wanted to make the impact of a possible dependence relevant to such researchers. The first data source is the GBA (Gemeentelijke BasisAdministratie) which is the official Dutch Population Register containing demographic information on the de jure population. The de jure population differs from the de facto population, the latter also containing residents who immigrated from other countries of the European Union and did not register as such, immigrants who (are planning to) stay shorter than four months and illegal immigrants. An important part of the difference between the de jure and the de facto population is the group of temporary workers from eastern Europe, in particular Poland. The second datasource is the HKS (HerKenningsdienst Systeem), which is a police register of all suspects of known offenses. We refer the reader to Van der Heijden et al. (2012) for more details on the registers.

2.2 Two registers without covariates

The simplest population size estimation model makes use of two registers, 1 and 2. Let variables A and B respectively denote inclusion in registers 1 and 2. Let the levels of A be indexed by i (i = 0,1) where $i = 0$ stands for "not included in register 1", and $i = 1$, stands for "included in register 1". Similarly, let the levels of B be indexed by j ($j = 0, 1$). Expected values are denoted by m_{ij}. Observed values are denoted by n_{ij} with $n_{00} = 0$, because there are no observations for the cases that belong to the population but were not present in either of the registers.

Recall that one of the assumptions in population size estimation is that the probability of being in the first register is independent of the probability of being in the second register. Under independence the loglinear model for the counts n_{01}, n_{10} and n_{11} is:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B. \tag{2.1}$$

where we used the identifying restrictions $\lambda_0^A = \lambda_0^B = 0$. There are two ways to derive the estimate of the missed part of the population. First, by $\hat{m}_{00} = \exp(\hat{\lambda})$ and second, by using the property that the odds ratio under independence is 1, i.e., $m_{00}m_{11}/m_{10}m_{01} = 1$ so that:

$$\hat{m}_{00} = \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10}n_{01}}{n_{11}}. \tag{2.2}$$

For the first way of estimating the missed portion of the population we need an estimate of λ in (3.1). There are several ways to estimate the parameters in (3.1), and it suits our purposes later on to use the generalized linear model. We assume that n_{ij} follow a Poisson distribution; a log link connects the expected values m_{ij} to the linear predictor. In terms of matrices and vectors we get

$$\log \begin{pmatrix} m_{11} \\ m_{10} \\ m_{01} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^A \\ \lambda_1^B \end{pmatrix} \tag{2.3}$$

where the right-hand side of (2.3) leads to a vector with elements $[\lambda + \lambda_1^A + \lambda_1^B, \lambda + \lambda_1^A, \lambda + \lambda_1^B]$. Thus the estimates of λ, λ_1^A and λ_1^B will get us estimates $\hat{m}_{11}, \hat{m}_{10}$ and $\hat{m}_{01}$ of which also the missed portion of the population $\hat{m}_{00}$ is found by $\log(\hat{m}_{00}) = \hat{\lambda}$, so that $\hat{m}_{00} = \exp(\hat{\lambda})$.

However, the problem with using the independence model is that independence is an unverifiable assumption, that is, we can not verify independence from the data. Thus the Poisson loglinear model for independence works under the assumption that the interaction parameter $\lambda_{ij}^{AB} = 0$. As noted before, this assumption could be violated and the population size estimate under independence may well be inaccurate. We are interested in what happens to the population size estimate when we assume independence when actually the inclusion probabilities of inclusion in registers A and B are dependent.

The approach we advocate is to include a fixed interaction parameter $\widetilde{\lambda}_{ij}^{AB}$ in the model, where tilde indicates that the interaction parameter is not estimated but fixed. By choosing interesting values for $\widetilde{\lambda}_{ij}^{AB}$ we can conduct a sensitivity analysis on the population size estimate. The loglinear model then becomes:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \widetilde{\lambda}_{ij}^{AB}. \tag{2.4}$$

where we used the identifying restrictions $\widetilde{\lambda}_{00}^{AB} = \widetilde{\lambda}_{10}^{AB} = \widetilde{\lambda}_{01}^{AB} = 0$. In matrix terms we get:

$$\log \begin{pmatrix} m_{11} \\ m_{10} \\ m_{01} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^A \\ \lambda_1^B \\ \widetilde{\lambda}_{11}^{AB} \end{pmatrix} \tag{2.5}$$

The loglinear model for independence is a special case of this saturated model when $\lambda_{ij}^{AB} = \widetilde{\lambda}_{ij}^{AB} = 0$. Dependence can be introduced to loglinear models by fixing $\widetilde{\lambda}_{ij}^{AB}$ to anything but 0. In software for Poisson regression, model (2.4) and (2.5) can be fit by entering $\widetilde{\lambda}_{ij}^{AB}$ as a so called offset. When $\widetilde{\lambda}_{ij}^{AB} \neq 0$, $\hat{\lambda}$ in (2.5) differs from $\hat{\lambda}$ in (2.3).

Note that interesting values for $\widetilde{\lambda}_{ij}^{AB}$ can be chosen using a direct relationship between λ_{ij}^{AB} and the odds ratio θ, which is:

$$\theta = \frac{m_{11}m_{00}}{m_{10}m_{01}} = \exp \widetilde{\lambda}_{11}^{AB}. \tag{2.6}$$

Using the Poisson loglinear model with an offset is a general approach for carrying out a sensitivity analysis. The approach is general in the sense that it can be applied in more complicated loglinear models, for example when it is desirable to investigate violations of more than one assumption simultaneously (compare models discussed in section 4.2). For completeness we also discuss a second method that is simpler but less general.

The second way of estimating the missed portion of the population is in using odds ratios directly, as has been done in Brown et al. (1999). We show this second way to give a full overview of the method. Also this provides for simpler notation that we will use in the rest of the paper. Under independence the odds ratio $m_{11}m_{00}/m_{10}m_{01} = 1$, and by rewriting and replacing the expected values with observed values, we get maximum likelihood estimate (3.2). We can impute dependence by making the odds ratio $\theta \neq 1$. Thus $\theta = m_{11}m_{00}/m_{10}m_{01}$, and

$$\hat{m}_{00(\theta)} = \theta \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \theta \frac{n_{10}n_{01}}{n_{11}} = \theta \hat{m}_{00}. \tag{2.7}$$

Note that $\hat{m}_{00(\theta)}$ can be found simply by multiplying the estimate under independence, $\hat{m}_{00}$, with θ. Both approaches, the loglinear Poisson regression with an offset and the odds ratio, yield the same $\hat{m}_{00}$. We will use the odds ratio to denote dependence as it provides a simpler notation than the interaction parameter $\widetilde{\lambda}_{ij}^{AB}$.

The methods just described allow us to study the impact of a violation of the independence assumption as a function of θ. To get the population size estimate, let n be the total of observed cases, $n = n_{01} + n_{10} + n_{11}$, let $\hat{N}$ be the population size estimated under $\theta = 1$, thus $\hat{N} = n + \hat{m}_{00}$, and define $\hat{N}_{(\theta)}$ as the estimated population size under dependence of size θ, $\hat{N}_{(\theta)} = n + \hat{m}_{00(\theta)} = n + \theta\hat{m}_{00}$. It follows that under negative dependence (i.e. $\theta < 1$), $\hat{N}$ will be an overestimation compared to $\hat{N}_{(\theta)}$ and under a positive dependence (i.e. $\theta > 1$), $\hat{N}$ will be an underestimation compared to $\hat{N}_{(\theta)}$. The bias will be smaller the closer θ is to 1.

Assume that A has a better coverage of the population than B. Then, when $n_{11}/(n_{11} + n_{01})$ is high the observed coverage is high, and vice versa. Brown et al. (2006) showed that as the observed coverage increases, the number of individuals that are missed by A reduces and $n_{11}/n_{10}n_{01}$ increases so that $n_{10}n_{01}/n_{11} = \hat{m}_{00}$ decreases. Then, the implied coverage of A is high, so that $\hat{m}_{00}$ is reasonably robust to dependence. When the observed coverage decreases, the number of individuals missed by A increases and $n_{11}/n_{10}n_{01}$ decreases. Then the implied coverage of A will be low, so that $\hat{m}_{00}$ is less robust to dependence.

To illustrate, we use two registers of Statistics Netherlands, the GBA and the HKS, on people with an Afghan, Iranian or Iraqi (AII) nationality living in the Netherlands in 2007 (shown in Table 3.2; Van der Heijden et al., 2012), and on people with a Polish nationality living in the Netherlands in 2009 (shown in Table 3.2; Van der Heijden et al., 2011).

TABLE 2.1
The observed values for the two nationalities, with the Afghan, Iraqi and Iranian people residing in the Netherlands in 2007 on the left, and the Polish people residing in the Netherlands in 2009 on the right.

	HKS	
GBA	1	0
1	1,085	26,254
0	255	-

	HKS	
GBA	1	0
1	374	39,488
0	1,445	-

For the people with an Afghan, Iraqi and Iranian nationality $\hat{m}_{00} = 6,170$ under independence between A and B. The population size estimated under $\theta = 1$ becomes $\hat{N} = 27,594 + 6,170 = 33,764$. Then, under dependence between A and B the estimated population size becomes $\hat{N}_{(\theta)} = 27,594 + (\theta * 6,170)$, see (2.7).

To investigate the robustness of the estimate under dependence we vary θ from 0.5 to 2. In the loglinear Poisson regression approach this corresponds to using offsets varying between log(0.5) and log(2). Table 2.2 shows $\hat{m}_{00(\theta)}$, the population size estimate $\hat{N}_{(\theta)}$, the estimated relative bias $\hat{N}/\hat{N}_{(\theta)}$ and the bootstrapped standard error (se) of the estimate for both nationalities (details about the parametric bootstrap are provided in section 5.3). As can be seen from the upper panel of Table 2.2, for the people with an Afghan, Iraqi and Iranian nationality under a dependence of $\theta = 0.5$, the estimate $\hat{m}_{00(\theta)}$ is half the size of the population size estimate under independence, and for a dependence of $\theta = 2$ the estimate $\hat{m}_{00}$ is twice the size of the population size estimate under independence. If in the population the registers are dependent with a true size θ, the population size estimate under independence varies between a 10 percent overestimation and a 15 percent underestimation. Thus when the true $\theta \neq 1$ our population size estimate under independence remains fairly accurate.

However, for the Polish people the population size estimate under dependence is not robust. As can be seen from the lower panel of Table 2.2, if in the population the registers are dependent with a true size θ, the population size estimate under independence deviates between a 65 percent overestimation and 44 percent underestimation. Thus when the true $\theta \neq 1$ the population size estimate under independence for the Polish people is not robust.

The most important reason that the population size estimate deviates this much is because the implied coverage of the people with an Afghan, Iraqi and Iranian nationality is smaller than for the individuals with a Polish nationality. For example, 1,085 is $1,085/(1,085 + 255) = 0.81$, thus 81 percent of implied coverage of the GBA measured by the HKS. Whereas for the individuals with a Polish nationality the implied coverage of the GBA is only 21 percent, confirming the research by Brown et al. (2006) that as the observed coverage increases the implied coverage increases and thus the population size estimate is more robust against dependence.

The estimated standard error of $\hat{N}_{(\theta)}$ is mainly determined by the size of $\hat{m}_{00(\theta)}$, and this explains the sharp rise of the standard error from $\theta = .50$ to $\theta = 2.00$ and the difference in standard error between the individuals with an Afghan, Iraqi and Iranian nationality and the individuals with a Polish nationality.

TABLE 2.2
Sensitivity analysis of the population size estimate for the people residing in the Netherlands in 2007 with an Afghan, Iraqi and Iranian nationality (upper panel) and for people with a Polish nationality in 2009 (lower panel).

		Odds Ratio				
		0.50	0.67	1.00	1.50	2.00
AII	$\hat{m}_{00(\theta)}$	3,085	4,114	6,170	9,255	12,341
	$\hat{N}_{(\theta)}$	30,679	31,708	33,764	36,849	39,935
	$\hat{N}/\hat{N}_{(\theta)}$	1.10	1.06	1.00	0.92	0.85
	Se	223	293	441	647	864
Polish	$\hat{m}_{00(\theta)}$	76,284	101,712	152,567	228,851	305,135
	$\hat{N}_{(\theta)}$	117,591	143,019	193,874	270,158	346,442
	$\hat{N}/\hat{N}_{(\theta)}$	1.65	1.36	1.00	0.72	0.56
	Se	4,473	6,024	8,787	13,630	17,866

2.3 Two registers with fully observed covariates

Covariates were first introduced to capture-recapture by Alho (1990) to reduce the heterogeneity resulting from individual differences on that covariate. As such, if covariates are available, the generally non-feasible independence assumption can be replaced with a less strict conditional independence assumption, where independence is conditional on covariates (Bishop et al., 1975; Van der Heijden et al., 2012). This assumption is less stringent because it can take into account inclusion probabilities that are heterogeneous over the levels of the included covariate. Another advantage of using covariates is that it allows us to investigate the characteristics of the missing portion of the population.

Suppose we have observed covariate X, where the levels of X are indexed by x, $(x = 0, 1)$. Under independence conditional on X there are two zero counts for cases not found in either register, namely for $x = 0$ and for $x = 1$. Let m_{ijx} denote the expected values for A, B and X. The loglinear model for independence for two registers and covariate X is:

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX}, \tag{2.8}$$

with identifying restrictions that a parameter equals zero when i or j or $x = 0$. When assuming independence between A and B conditional on X, $\lambda_{ij}^{AB} = \lambda_{ijx}^{ABX} = 0$. We use the notation of Bishop et al. (1975) to denote hierarchical loglinear models, i.e. we denote this model as $[AX][BX]$.

In section 2 we discussed two ways to estimate population sizes in a sensitivity analysis, namely one using an offset in a Poisson loglinear model and another using odds ratios directly. Here we only discuss the first way as it is more general. We assume that n_{ijx} follow a Poisson distribution and a log link connects the expected value m_{ijx} to the linear predictor.

It is important to note that also in this context sensitivity analyses are useful for assessing the impact of assumptions that are not verifiable from the data under study. Here conditional independence is the unverifiable assumption since model $[AX][BX]$ is the saturated model. In contrast, model violations for more restricted models are verifiable in the data, for example for a model such as $[A][BX]$. Hence, the impact of interaction between A and X does not have to be investigated via a sensitivity analysis. However, when there may be dependence between A and B a sensitivity analysis is useful.

We model dependence in the data by adding fixed parameters $\widetilde{\lambda}_{ij}^{AB} + \widetilde{\lambda}_{ijx}^{ABX}$ to model 2.8. We again work under the saturated model, as the number of parameters to be estimated is equal to the number of observed parameters:

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX} + \widetilde{\lambda}_{ij}^{AB} + \widetilde{\lambda}_{ijx}^{ABX}, \quad (2.9)$$

with the additional restrictions that parameters $\widetilde{\lambda}_{ij}^{AB}$ and $\widetilde{\lambda}_{ijx}^{ABX}$ equal zero when i or j or $x = 0$.

Under dependence between A and B given X, the association between the odds ratio θ_x and the loglinear parameters is:

$$\theta_x = \frac{m_{11x}m_{00x}}{m_{10x}m_{01x}} = \exp(\widetilde{\lambda}_{11}^{AB} + \widetilde{\lambda}_{11x}^{ABX}). \quad (2.10)$$

When we assume that dependence for $x = 0$ is identical to dependence for $x = 1$, then:

$$\theta = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp(\widetilde{\lambda}_{11}^{AB}). \quad (2.11)$$

We estimate (2.9) using loglinear Poisson regression with for cell (1,1,0) the offset $\widetilde{\lambda}_{11}^{AB}$ and for cell (1,1,1) the offset $\widetilde{\lambda}_{11}^{AB} + \widetilde{\lambda}_{111}^{ABX}$. After estimating (2.9), estimates for the missed portions of the population are found by $\hat{m}_{000} = \exp(\hat{\lambda})$ and $\hat{m}_{001} = \exp\ (\hat{\lambda} + \hat{\lambda}_1^X)$.

Table 2.3 shows the data for the Afghan, Iraqi and Iranian people distributed over males ($x = 0$) and females ($x = 1$). Under conditional independence $\hat{m}_{000} = 3,583$ and $\hat{m}_{001} = 2,113$. Taken together both registers missed 5,696 cases. Note that, conditional independence does not imply marginal independence under model $[AX][BX]$, since the marginal odds ratio $1,085 * 5,696/26,254 * 255 = 0.92$, and hence shows dependence (under marginal independence it would be equal to 1).

TABLE 2.3
The observed values for the Afghan, Iraqi and Iranian people, on the left panel the males and on the right panel the females.

	HKS	
GBA	1	0
1	972	14,883
0	234	-

	HKS	
GBA	1	0
1	113	11,371
0	21	-

We estimate the parameters in (2.9) with a Poisson regression with $\widetilde{\lambda}_{ijx}^{ABX} = 0$, so that the odds ratio of the males equals the odds ratio of the females (compare (2.11)). The upper panel of Table 2.5 shows the results of the sensitivity analysis for the people with an Afghan, Iraqi and Iranian nationality in 2007 and the covariate gender. If in the population the registers are dependent with a true size θ, the population

TABLE 2.4
The observed values for the Polish people, on the left panel the males and on the right panel the females.

	HKS	
GBA	1	0
1	313	19,152
0	1,349	-

	HKS	
GBA	1	0
1	61	20,336
0	96	-

size estimate under independence varies between a 9 percent overestimation to a 15 percent underestimation. As $\hat{m}_{00(\theta)}$ is relatively small, the standard error is relatively small. Thus when the true $\theta = 0.5$ but we estimate under $\theta = 1$ the population size estimate under independence is fairly robust.

For the people residing in the Netherlands with a Polish nationality in 2009 the covariate gender is also used. Under conditional independence the estimate $\hat{m}_{00x} = 144,548$. The lower panel of Table 2.5 shows the sensitivity analysis of the population size estimator under conditional independence. If in the population the registers are dependent with a true size θ, the population size estimate under independence ranged between a 58 percent overestimation and a 42 percent underestimation. Thus when the true $\theta \neq 1$ the population size estimate deviates greatly from the population size estimate under $\theta = 1$, indicating that for this dataset the population size estimate under independence is not robust.

TABLE 2.5
Sensitivity analysis for the people with an Afghan, Iraqi and Iranian (AII) nationality residing in the Netherlands in 2007 (upper panel), and the people with a Polish nationality residing in the Netherlands in 2009 (lower panel), conditional on gender.

		Odds Ratio				
		0.50	0.67	1.00	1.50	2.00
AII	$\hat{m}_{00}$	2,848	3,797	5,696	8,544	11,392
	$\hat{N}_{(\theta)}$	30,442	31,391	33,290	36,138	38,986
	$\hat{N}/\hat{N}_{(\theta)}$	1.09	1.06	1.00	0.92	0.85
	Se	292	390	576	863	1144
Polish	$\hat{m}_{00}$	57,274	76,365	114,548	171,821	229,095
	$\hat{N}_{(\theta)}$	98,581	117,672	155,855	213,128	270,402
	$\hat{N}/\hat{N}_{(\theta)}$	1.58	1.32	1.00	0.73	0.58
	Se	3,814	5,088	7450	11,465	15,135

We note that this example is using a covariate with only two levels. One can easily extend this to covariates with more levels. Assume covariate W has three levels, where the levels of W are indexed by w, $(w = 0, 1, 2)$. Then there are three zero counts, namely for $w = 0$, $w = 1$ and $w = 2$. One can estimate the zero counts using equation (2.10), where estimates for the missed portions of the population are found by $\hat{m}_{000} = \exp(\hat{\lambda})$ and $\hat{m}_{001} = \exp\ (\hat{\lambda} + \hat{\lambda}_1^W)$ and $\hat{m}_{002} = \exp\ (\hat{\lambda} + \hat{\lambda}_2^W)$.

2.4 Two registers with partially observed covariates

In section 3 we have used covariates that are present in both registers (fully observed covariates) to replace the strict independence assumption with an independence assumption conditional on covariates. However, a register usually also has a set of variables that are only measured in one register and not in the other register (partially observed covariates). Partially observed covariates in A are usually ignored because including them leads to missing data in B for those individuals that are not in A, and vice versa. When these covariates are related to the inclusion probability, ignoring the partially observed covariates can lead to a biased population size estimate (Zwane and Van der Heijden, 2007; Van der Heijden et al., 2012).

2.4.1 Partially observed covariates

Partially observed covariates can be approached as a missing data problem (Zwane and Van der Heijden, 2007). If we assume a Missing At Random (MAR) mechanism for the data, then we can use the Expectation-Maximization (EM) algorithm to estimate the missing values of the partially observed covariate of register A (and B) for the individuals not present in A (and B). MAR assumes that the probability of missingness depends only on the observed variables in the capture-recapture model (Little and Rubin, 2002). When the assumption of MAR has been satisfied the EM algorithm will give unbiased estimates.

Suppose register A has the covariate X_1, indexed by $k(k = 0, 1)$, where the values for X_1 are missing for $A = 0$ because X_1 is not in register B. Assume that register B has the covariate X_2, indexed by $l(l = 0, 1)$, where the values for X_2 are missing for $B = 0$ because X_2 is not in register A. The loglinear conditional independence model for two registers, with two partially observed covariates X_1 and X_2, is denoted as

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2}, \quad (2.12)$$

with identifying restrictions $\lambda_{ij}^{AB} = \lambda_{ik}^{AX_1} = \lambda_{jl}^{BX_2} = \lambda_{ijk}^{ABX_1} = \lambda_{ijl}^{ABX_2} = \lambda_{ijkl}^{ABX_1X_2} = 0$. The conditional independence model is denoted by $[AX_2][BX_1][X_1X_2]$. Inclusion of the parameter $\lambda_{il}^{AX_2}$ instead of the parameter $\lambda_{ik}^{AX_1}$ may seem counterintuitive but an interaction for A and X_1 cannot be identified as the levels of X_1 do not vary over individuals for which $A = 0$, and similarly for B and X_2 (Zwane and Van der Heijden, 2007).

Table 2.6 illustrates that two registers with two covariates lead to

TABLE 2.6
Expected values for two registers and two partially observed covariates.

		B = 1		B = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
A = 1	$X_1 = 1$	m_{1111}	m_{1110}	m_{1011}	m_{1010}
	$X_1 = 0$	m_{1101}	m_{1100}	m_{1001}	m_{1000}
A = 0	$X_1 = 1$	m_{0111}	m_{0110}	m_{0011}	m_{0010}
	$X_1 = 0$	m_{0101}	m_{0100}	m_{0001}	m_{0000}

16 cells. However, because our covariates are only partially observed, columns $X_2 = 1$ and $X_2 = 0$ for $B = 0$ are collapsed, just as rows $X_1 = 1$ and $X_1 = 0$ for $A = 0$ are collapsed. In other words, we do not observe counts for m_{0111} and m_{0101} but only one count for the sum $m_{0111}+m_{0101}$, and similarly for $m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$ and $m_{1001} + m_{1000}$. Note that we have no observed values for m_{0011}, m_{0001}, m_{0010} and m_{0000}, as these refer to individuals who are in neither of the registers. Thus model $[AX_2][BX_1][X_1X_2]$ is saturated with eight observed values and eight parameters to be estimated.

Using the EM algorithm we first estimate the four missing cells, i.e. the cells that are missing because the covariates are only partially observed. In the E-step we spread out the four sums $m_{0111} + m_{0101}$, $m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$ and $m_{1001} + m_{1000}$ over the eight cells to get an expectation for the missing data. In the M-step we estimate loglinear model (2.12) to the completed table of 12 cells. For estimation we assume that the 12 counts follow a Poisson distribution and a log link connects the expected counts to the linear predictor. The resulting estimates are then used for the E-step where in the M-step, following (2.12), we estimate the parameters again.

To illustrate we use again the data on the people with an Afghan, Iraqi and Iranian nationality residing in the Netherlands in 2007 with two partially observed covariates (Van der Heijden et al., 2012). The GBA has the partially observed covariate marital status (X_1), where $X_1 = 1$ denotes either being married or living together and $X_1 = 0$ denotes either unmarried, divorced or widowed. The HKS has the partially observed covariate police region (X_2), where $X_2 = 1$ denotes residing in one of the 5 biggest cities of the Netherlands (i.e. Amsterdam, Rotterdam, Utrecht, The Hague and Eindhoven) and $X_2 = 0$ denotes residing in the rest of the country.

Due to the loglinear model used the first four observed values remain unchanged for each iteration (for $GBA = 1$ and $HKS = 1$). The upper panel of Table 2.7 shows the observed counts and the lower panel of Table 2.7 shows the fitted counts after convergence of the EM-algorithm. As an example, the observed value of 91 (for $X_2 = 1$, where X_1 values are missing under $GBA = 0$) is spread out into the values 64 for $X_1 = 1$ and 27 for $X_1 = 0$. After convergence the unobserved part of the population is estimated. In total, we estimate that there were 33,770 individuals with an Afghan, Iraqi and Iranian nationality residing in the Netherlands in 2007.

TABLE 2.7
Data for the Afghan, Iraqi and Iranian people residing in the Netherlands in 2007, spread out over the partially observed covariates Marital status X_1 and Police region X_2

Panel 1: The observed counts

		HKS = 1		HKS = 0
		$X_2 = 1$	$X_2 = 0$	X_2 missing
GBA = 1	$X_1 = 1$	259	539	13,898
	$X_1 = 0$	110	177	12,356
GBA = 0	X_1 missing	91	164	-

Panel 2: The fitted frequencies

		HKS = 1		HKS = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
GBA = 1	$X_1 = 1$	259	539	4,511	9,387
	$X_1 = 0$	110	177	4,736	7,620
GBA = 0	$X_1 = 1$	64	123	1,112	2,150
	$X_1 = 0$	27	41	1,168	1,745

2.4.2 Sensitivity analyses

We again make use of a sensitivity analysis to investigate the unverifiable assumption of independence conditional on partially observed covariates. Model violations for more restricted models are verifiable in the data. For example, using a model such as $[AX_2][BX_1]$ allows us to investigate absence of interaction $\lambda_{kl}^{X_1X_2}$ in the data. Thus the impact of an interaction between X_1 and X_2 does not need to be investigated via a sensitivity analysis. However in this context (2.12) is the saturated model and therefore model violations such as dependence between A and X_1, between B and X_2, and between A and B are unverifiable, rendering it useful to conduct a sensitivity analysis. Note that in the previous paragraphs we have used a sensitivity analysis to assess the interaction between the two registers. In this paragraph we assess not only the interaction between A and B, but also the interaction between the register and its partially observed covariate. To exemplify, we introduce an interaction parameter that simulates dependence between the GBA and marital status. Such a dependence would imply that marital status influences the inclusion probability of being in the GBA.

The loglinear model for an interaction between A and B would be:

$$\log m_{ijkl} = \lambda+\lambda_i^A+\lambda_j^B+\lambda_k^{X_1}+\lambda_l^{X_2}+\lambda_{il}^{AX_2}+\lambda_{jk}^{BX_1}+\lambda_{kl}^{X_1X_2}+\widetilde{\lambda}_{ij}^{AB}, \quad (2.13)$$

with additional identifying restrictions that $\widetilde{\lambda}_{ij}^{AB} = 0$ when i or j equals 0. Here $\exp(\widetilde{\lambda}_{ij}^{AB})$ is the conditional odds ratio for the interaction between A and B.

Assume the partially observed covariate marital status is related to the inclusion probability of the GBA, thus $\lambda_{ik}^{AX_1} \neq 0$. Because the interaction between A and X_1 is unverifiable from the data the fixed parameter $\widetilde{\lambda}_{ik}^{AX_1}$ has been added to the loglinear model (2.12). We continue to work under the saturated model:

$$\log m_{ijkl} = \lambda+\lambda_i^A+\lambda_j^B+\lambda_k^{X_1}+\lambda_l^{X_2}+\lambda_{il}^{AX_2}+\lambda_{jk}^{BX_1}+\lambda_{kl}^{X_1X_2}+\widetilde{\lambda}_{ik}^{AX_1}, \quad (2.14)$$

with additional identifying restrictions that $\widetilde{\lambda}_{ik}^{AX_1} = 0$ when i or k equals 0. The same can be done for the interaction between B and X_2. When the partially observed covariate X_2 is related to the inclusion probability of register B, $\lambda_{jl}^{BX_2} \neq 0$. We add fixed parameter $\widetilde{\lambda}_{jl}^{BX_2}$ to the loglinear model. The loglinear model then becomes:

$$\log m_{ijkl} = \lambda+\lambda_i^A+\lambda_j^B+\lambda_k^{X_1}+\lambda_l^{X_2}+\lambda_{il}^{AX_2}+\lambda_{jk}^{BX_1}+\lambda_{kl}^{X_1X_2}+\widetilde{\lambda}_{jl}^{BX_2}, \quad (2.15)$$

with additional identifying restrictions that $\widetilde{\lambda}_{jl}^{BX_2} = 0$ when j or l equals 0. We can estimate (2.13), (2.14) and (2.15) via Poisson regressions

TABLE 2.8
Sensitivity analysis of the population size estimate for the people residing in the Netherlands in 2007 with an Afghan, Iraqi and Iranian nationality with the interaction A and X_1 (upper panel) and the interaction between B and X_2 (lower panel).

		Odds Ratio				
		0.50	0.67	1.00	1.50	2.00
AB	$\hat{m}_{00(\theta)}$	3.088	4,117	6,176	9,264	12,352
	$\hat{N}_{(\theta)}$	30.682	31,711	33,770	36,858	39,946
	$\hat{N}/\hat{N}_{(\theta)}$	1.10	1.06	1.00	0.92	0.85
AX1	$\hat{m}_{00(\theta)}$	5,443	5,711	6,176	6,736	7,179
	$\hat{N}_{(\theta)}$	33,037	33,305	33,770	34,330	34,773
	$\hat{N}/\hat{N}_{(\theta)}$	1.0222	1.0140	1.00	0.9837	0.9711
BX2	$\hat{m}_{00(\theta)}$	6,253	6,220	6,176	6,136	6,112
	$\hat{N}_{(\theta)}$	33,847	33,814	33,770	33,730	33,706
	$\hat{N}/\hat{N}_{(\theta)}$	0.9977	0.9987	1.00	1.0012	1.0019

with offsets. Do note that in modeling these relationships we have to fix the offset variable on a log scale. Then we can estimate the portions of the population that both registers have missed by $\hat{m}_{0000} = \exp(\hat{\lambda})$, $\hat{m}_{0010} = \exp(\hat{\lambda} + \hat{\lambda}_1^{X_1})$, $\hat{m}_{0001} = \exp(\hat{\lambda} + \hat{\lambda}_1^{X_2})$ and $\hat{m}_{0011} = \exp(\hat{\lambda} + \hat{\lambda}_1^{X_1} + \hat{\lambda}_1^{X_2} + \hat{\lambda}_{11}^{X_1X_2})$.

The upper panel of Table 2.8 shows the sensitivity analysis for the interaction between A and B, the middle panel shows the sensitivity analysis for the interaction between A and X_1 and the lower panel shows the sensitivity analysis for the interaction between B and X_2 for the Afghan, Iraqi and Iranian people. As can be seen, for the interaction between A and B, the relative bias is similar to the bias found in Tables 2.2 and 2.5. If in the population the GBA and marital status are dependent with a true size θ, the estimation under independence deviates between a 2.22 percent overestimation to a 2.89 percent underestimation, and the estimation under independence between the HKS and police region deviates between a 0.23 percent underestimation and a 0.19 percent overestimation. Thus for the interactions AX_1 and BX_2, when the true $\theta \neq 1$, the population size estimate under independence remains fairly robust.

We have done the same for the people with a Polish nationality residing in the Netherlands in 2009. The observed values are shown in the upper panel of Table 2.9 and the expected frequencies are shown in the lower panel of Table 2.9. Again a sensitivity analysis has been conducted, which is shown in Table 2.10. Just as with the individuals with an Afghan, Iraqi and Iranian nationality the estimates and thus the relative bias under dependence between A and B remains unchanged (compare Tables (2.2) and (2.5)). If in the population the GBA and marital status are dependent with a true size θ, the population size estimate under independence ranges from a 7 percent overestimation to a 9 percent underestimation (upper panel). The estimate under independence between the HKS and police region deviates from a 2 percent underestimation to a 2 percent overestimation (lower panel). Thus when the true $\theta \neq 1$, the population size estimate under independence remains fairly robust.

Under the use of partially observed covariates it becomes clear why the loglinear Poisson regression provides a more general approach than using odds ratios to implement the sensitivity analyses. In using loglinear Poisson regression the process becomes vastly simpler, in that the offset can be set to any number per cell. When multiple different offsets are in use, the loglinear Poisson regression allows for this complexity, whereas implementing odds ratios may become gruesome.

TABLE 2.9
The observed counts for the people with a Polish nationality residing in the Netherlands in 2009 (upper panel) and the fitted frequencies spread out over the partially observed covariates (lower panel).

Panel 1: The observed counts

		HKS = 1		HKS = 0
		$X_2 = 1$	$X_2 = 0$	X_2 missing
GBA = 1	$X_1 = 1$	111	188	25,416
	$X_1 = 2$	32	43	14,072
GBA = 0	$X_1 = 1$	603	842	

Panel 2: The fitted frequencies

		HKS = 1		HKS = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
GBA = 1	$X_1 = 1$	111	188	9,435	15,981
	$X_1 = 2$	32	43	6,004	8,068
GBA = 0	$X_1 = 1$	468	685	39,787	58,250
	$X_1 = 2$	135	157	25,318	29,408

TABLE 2.10
Sensitivity analysis of the population size estimate for the the people residing in the Netherlands in 2009 with a Polish nationality with the interaction between A and X_1 (upper panel) and the interaction between B and X_2 (lower panel).

		Odds Ratio				
		0.50	0.67	1.00	1.50	2.00
AB	$\hat{m}_{00(\theta)}$	76,381	101,842	152,762	229,143	305,524
	$\hat{N}_{(\theta)}$	117,688	143,149	194,069	270,450	346,832
	$\hat{N}/\hat{N}_{(\theta)}$	1.65	1.36	1.00	0.71	0.56
AX1	$\hat{m}_{00(\theta)}$	139,494	144,238	152,762	163,584	172,582
	$\hat{N}_{(\theta)}$	180,801	185,545	194,069	204,891	213,889
	$\hat{N}/\hat{N}_{(\theta)}$	1.07	1.05	1.00	0.95	0.91
BX2	$\hat{m}_{00(\theta)}$	156,616	155,004	152,762	150,707	149,429
	$\hat{N}_{(\theta)}$	197,923	196,311	194,069	192,014	190,736
	$\hat{N}/\hat{N}_{(\theta)}$	0.98	0.99	1.00	1.01	1.02

2.5 Miscellany

2.5.1 Extension to multiple sources

One way to make the impact of possible violations of the independence assumption less severe is by conditioning on covariates, as we have seen in section 3 and 4. Another way to make the impact of possible violations of the independence assumption less severe is by adding registers, when more registers are available (compare, Baffour et al., 2013). Assume we have three registers 1, 2 and 3, where respectively variables ’A’, ’B’ and ’C’ stand for inclusion in the registers. We denote the expected values m_{ijp} where $i, j, p = 1$ stand for the inclusion into A, B and C respectively and where i, j, $p = 0$ stands for the absence in A, B and C.

For three variables the saturated loglinear model is denoted by:

$$\log m_{ijp} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_p^C + \lambda_{ij}^{AB} + \lambda_{ip}^{AC} + \lambda_{jp}^{BC}, \tag{2.16}$$

with identifying restrictions that a parameter equals zero when i, j or $p = 0$. We assume that interaction parameter $\lambda_{ijp}^{ABC} = 0$. Model $[AB][BC][AC]$ is the saturated model, as the number of observed parameters equals the number of parameters to be estimated. With d registers, we assume that the d-factor interaction is absent.

For estimation assume that n_{ijp} follow a Poisson distribution and a log link connects the expected value m_{ijp} to the linear predictor. We can estimate the parameters in (5.1) via a Poisson loglinear regression.

Model $[AB][BC][AC]$ assumes that odds conditional on a third variable are equal, for example for the odds ratio between A and B given C we find

$$\frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}}. \tag{2.17}$$

Model (5.1) assumes that for estimation with odds ratios under saturated model $[AB][BC][AC]$ we get:

$$\frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}} = \hat{m}_{000}. \tag{2.18}$$

An estimate for $\hat{m}_{000}$ is easily derived from (2.17) as $[AB][AC][BC]$ is the saturated model in this context, absence of the three factor interaction is an unverifiable assumption as it can not be verified in the data. More restricted models such as $[AB][AC]$ are verifiable in the data. However, we can investigate the robustness of the population size estimate against violations of the assumption that the three factor interaction is absent by fixing the interaction parameter to anything but 0, i.e. $\widetilde{\lambda}_{ijp}^{ABC} \neq 0$. Thus the loglinear model becomes:

$$\log m_{ijp} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_p^C + \lambda_{ij}^{AB} + \lambda_{ip}^{AC} + \lambda_{jp}^{BC} + \widetilde{\lambda}_{ijp}^{ABC}, \tag{2.19}$$

with the additional identifying restriction where parameter $\widetilde{\lambda}_{ijp}^{ABC}$ equals zero when i or j or $p = 0$. The population size estimate under (2.19) can be estimated using Poisson loglinear regression with parameter $\widetilde{\lambda}_{ijp}^{ABC}$ as an offset.

Under dependence between A and B given C, the association between the odds ratio θ and the loglinear parameters is:

$$\theta_{AB}^{(p=0)} = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \exp(\lambda_{11}^{AB}), \tag{2.20}$$

and:

$$\theta_{AB}^{(p=1)} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp(\lambda_{11}^{AB} + \lambda_{111}^{ABC}). \tag{2.21}$$

When we assume that the odds ratio between A and B is the same for $p = 0$ and $p = 1$, we get

$$\theta_{AB} = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp(\lambda_{11}^{AB}). \tag{2.22}$$

When more registers are available we can use these extra registers to reduce the impact of violations of the independence assumption. As we have shown the loglinear model is easily generalizable to multiple registers.

2.5.2 Multiplier method

The multiplier method is an alternative method to estimate the size of a population and it is used, amongst others, in drug use research and HIV prevalence (European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), 1997; Cruts and Van Laar, 2010; Temurhan et al., 2011). Multiplier methods are user-friendly for their mathematical simplicity, the absence of linkage and are straightforward to use. At least two data sources are needed to use the multiplier method, usually a comprehensive register and a survey. For example, assume we wish to estimate the number of Polish people residing in the Netherlands in 2013. We assume that everyone has an equal chance of going to a hospital, thus we go to hospitals to assess how many Polish patients there are, and ask them whether they are in the GBA. Then assume the data we found is the data from Table 2.11. There are 200 Polish people, of which 150 are in the GBA. Thus $p(\text{GBA }|\text{Hospital}) = 0.75$. If in the GBA there is a total of 40,000 Polish people registered, this means our actual total should be $40,000/0.75 = 53,333$ and we missed $53,333 - 40,000 = 13,333$ people who are not registered in the GBA.

The multiplier method can also be explained from the perspective of capture recapture methods. Using the counts provided above we have n_{11}, n_{01} and n_{1+} so that $n_{1+} - n_{11} = n_{10}$ and equation (3.2) gives $(39,850 * 50)/150 = 13,283$. Then $\hat{N} = 150 + 50 + 39,850 + 13,283 = 53,333$, which is the exact same value as we got above. A sensitivity analysis could be conducted using equation (2.7).

The attractiveness of the multiplier method lies in the absence of linkage of two sources. When estimating hidden or hard-to-reach populations it is likely that it is difficult to get identifying variables to link the individuals in the samples. The absence of linkage is what makes the multiplier method different from capture-recapture. However, it has to be kept in mind that the multiplier method also relies on the underlying assumptions that being in the hospital is statistically independent from being in the GBA, and that it relies on individuals reporting their GBA status accurately when being admitted to the hospital.

TABLE 2.11
Artificial observed data for the Polish people in the hospital

		Hospital		
		1	0	
GBA	1	150	39,850	40,000
	0	50	-	-
		200	-	-

2.5.3 Confidence intervals

Apart from robustness, another aspect of the usefulness of a point estimate is its confidence interval. Parametric bootstrap confidence intervals can be used to find these confidence intervals in a simple way when dealing with incomplete contingency tables. In a parametric bootstrap sample the estimate $\hat{m}_{00(\theta)}$ for cell (0,0) is used in the multinomial probabilities. So for Table (3.2), the four probabilities are $n_{11}/\hat{N}_{(\theta)}$, $n_{10}/\hat{N}_{(\theta)}$, $n_{01}/\hat{N}_{(\theta)}$ and $\hat{m}_{00(\theta)}/\hat{N}_{(\theta)}$. A sample with size $\hat{N}_{(\theta)}$ is drawn with replacement. This yields four counts $n_{11}^{b=1}$, $n_{01}^{b=1}$, $n_{10}^{b=1}$ and $n_{00}^{b=1}$. The first bootstrap population size estimate $\hat{N}^{b=1}$ is found using only $n_{11}^{b=1}$, $n_{01}^{b=1}$, $n_{10}^{b=1}$, i.e. ignoring $n_{00}^{b=1}$, and estimating $\hat{m}_{00(\theta)}^{b=1}$. This is repeated 10,000 times, yielding 10,000 bootstrap population size estimates. From these 2.5 and 97.5 percentile scores are derived.

To exemplify we constructed a parametric bootstrapping confidence interval on the data presented in paragraph 2, which can be found in Table 2.12. The R code for the parametric bootstrap confidence interval can be found in Appendix A.3.

To compare, we also constructed asymptotic confidence estimate $CI = \hat{m}_{00} + / - z_{(.975)}(\sqrt{\hat{\text{Var}}(n)})$, where $\hat{\text{Var}}(n) = (n_{1+}n_{+1}n_{10}n_{01})/((n_{11})^3)$ (Bishop et al., 1975). The estimated confidence interval for the Afghan, Iraqi and Iranian people under independence is $32,905.44 - 34,623.16$, which is close to the bootstrapped confidence interval.

TABLE 2.12
Confidence intervals

Odds Ratio	AII	Polish
0.50	30,254 - 31,132	109,529 - 127,022
0.67	31,156 - 32,288	132,278 - 155,837
1.00	32,931 - 34,654	177,476 - 212,431
1.50	35,607 - 38,125	245,439 - 298,960
2.00	38,292 - 41,682	314,212 - 384,579

2.6 Discussion

We have shown for two different datasets that the population size estimate under dependence could be fairly robust as well as not robust at all.

Deviations from independence when implied coverage is low (and thus $\hat{m}_{00}$ is high) result in bigger deviations from the population size estimate under fixed dependence than when the implied coverage is higher. Thus the estimate becomes less robust and this makes the situation worse. For the Afghan, Iraqi and Iranian people the population size estimate did not change much when dependence was introduced; it also remained fairly robust whether or not we assumed conditional independence on fully observed covariates. However for the Polish people, the implied coverage is small resulting in a higher $\hat{m}_{00}$ so that the deviation from independence will be large. The resulting lack of robustness makes it even worse. Not only did the population size estimate under independence change dramatically under fixed dependence, adding a covariate to replace the strict independence assumption with the less strict independence assumption conditional on covariates changed the population size estimate but did not improve the robustness.

This reflects that the Polish people are, much more so than people from Afghanistan, Iraq and Iran, in the position that they work on a temporary basis without living permanently in The Netherlands. By law, it is approved for people from European Union countries like Poland to work without a working and living permit. This is not the case for people from Afghanistan, Iraq and Iran. Therefore, the coverage of the GBA differs between both nationalities which gives a relatively high estimation of the missed population of the Polish people compared to the Afghan, Iraqi and Iranian people. Additionally because we multiply $\hat{m}_{00}$ with θ it follows that a bigger $\hat{m}_{00}$ will result in a bigger $\hat{m}_{00\theta}$ than a smaller $\hat{m}_{00}$ would when multiplied with the same θ.

We also showed how to investigate robustness of the population size estimate in models with partially observed covariates. For the example we used the population size estimate was relatively insensitive to violation of specific conditional independence assumptions. Since adding covariates reduces heterogeneity and gives the opportunity to assess how the population is divided over the levels of the covariate it is useful to include a partially observed covariate.

In this manuscript we assumed that the only assumption that was violated was the independence assumption. However, also violation of other assumptions could have a large impact on the population size estimate. In particular, research on violation of the assumptions that the registers are perfectly linked as well as that the population is closed during the observation period is needed to conclude on the usefulness of the capture-recapture method for estimating the undercoverage of census data.

We have chosen a range of odds ratio from 0.5 to 2. As per our knowledge, it is not possible to get an accurate estimation of what a realistic θ value would be, since it is impossible to ascertain θ from the data. One way of dealing with the strict independence assumption is in adding a third register, hence using another source to estimate θ, as has been done by Brown et al. (2006) who created an adjustment factor based on a third source for the Census.

In conclusion, it is important to assess the size of the implied coverage of one of the registers. We have shown that lack of robustness under dependence is easily established when implied coverage is low. However, when implied coverage is high the population size estimate remains fairly robust. Thus, instead of accepting the population size estimate as is, researchers should report the robustness of their estimate.

2.7 Appendix

To estimate the population size under loglinear models we have used Poisson regression with an offset in SPSS and R.

2.7.1 R code

Below is given the R code to get estimates $\hat{m}_{00kl}$ in the EM algorithm, for the Polish data only.

```
##Give the data
data = c(111,188,32,43,12708,12708,7036,7036,301.5,421,301.5,421) ## Polish data
data = data*10000
freqitx = freqit1= data

    ## Design matrix
A = c(1,1,1,1,1,1,1,1,0,0,0,0)
B = c(1,1,1,1,0,0,0,0,1,1,1,1)
X1 = c(1,1,0,0,1,1,0,0,1,1,0,0)
X2 = c(1,0,1,0,1,0,1,0,1,0,1,0)

    ## OR for independence
offst=c(0,0,0,0,0,0,0,0,0,0,0,0)
for (i in 1:50000){
glm = glm(freqitx ~ A*X2+B*X1+X1*X2, offset=offst, family = poisson)
freqdata = c(data[1:4])
freqfit = glm$fitted.values[5:12]
freqitx=c(freqdata,freqfit)
freqitx=round(freqitx)}

    ## Parameter estimates under independence
par = glm$coefficients
m0011 = as.numeric(exp(par[1] + par[3] + par[5] + par[8]))
m0010 = as.numeric(exp(par[1] + par[5]))
m0001 = as.numeric(exp(par[1] + par[3]))
m0000 = as.numeric(exp(par[1]))
matrix = matrix(c(glm$fitted.values[1],glm$fitted.values[2],
glm$fitted.values[5],glm$fitted.values[6],glm$fitted.values[3],glm$fitted.values[4],
glm$fitted.values[7], glm$fitted.values[8], glm$fitted.values[9], glm$fitted.values[10],
m0011, m0010, glm$fitted.values[11],glm$fitted.values[12],m0001, m0000),
4,4, byrow=TRUE)
N = sum(matrix)

    ## Define the offsets. Here we only give an example for the offsets of BX2 = 0.5
offst1=c(-0.6931472,0,-0.6931472,0,0,0,0,0,-0.6931472,0,-0.6931472,0)
## Iterative GLM Loop for the EM algorithm
for (i in 1:50000){
glm = glm(freqitx ~ A*X2+B*X1+X1*X2, offset=offst1, family = poisson)
freqdata = c(data[1:4])
freqfit = glm$fitted.values[5:12]
freqitx=c(freqdata,freqfit)
freqitx=round(freqitx)}

    ## Calculation of estimated missed frequencies
par = glm$coefficients
m0011 = as.numeric(exp(par[1] + par[3] + par[5] + par[8]))
m0010 = as.numeric(exp(par[1] + par[5]))
m0001 = as.numeric(exp(par[1] + par[3]))
m0000 = as.numeric(exp(par[1]))

    m00comp = m0011+m0010+m0001+m0000
PSE = sum(data)+m00comp
print(m00comp)
print(sum(data)+m00comp)
print(N/PSE)
```

2.7.2 SPSS syntax

```
compute freqitx=freqit1.
compute freqitx=rnd(freqitx).
execute.
DEFINE EM_PGLM()
!DO !l = 1 !TO 10000.
GENLIN freqitx BY A B X1 X2 (ORDER=ASCENDING)
/MODEL A B X1 X2 A*X2 B*X1 X1*X2 INTERCEPT=YES OFFSET=offst05
DISTRIBUTION=POISSON LINK=LOG
/SAVE MEANPRED (pred_val).
compute diff=ABS(freqit1-pred_val).
means diff.
compute freqitx=pred_val.

IF((A = 1)&(B = 1)&(X1 = 1)&(X2 = 1))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 2)&(X2 = 1))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 1)&(X2 = 2))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 2)&(X2 = 2))freqitx = freqit1.

COMPUTE freqitx = rnd(freqitx).
execute.
delete variables pred_val.
!DOEND
!ENDDEFINE.
##run the macro
EM_PGLM.
```

2.7.3 R code parametic bootstrap

The R code presented below represents the parametric bootstrap for the Polish data from Table 3.2

```
data = c(374, 39488, 1445) ## Polish data
theta = 2
m00 = (data[2]*data[3])/data[1]
m00theta = m00*theta
datacomp = sum(data,m00theta)
## The estimate of N, under an offset theta
n = sum(data)
N = n + m00theta
##The relative bias under an offset theta
```

```
(n+m00)/N
## Parametric bootstrap
NN = c( N)
p = matrix(c(data/datacomp, m00theta/datacomp),1)
set.seed(N)
library(combinat)
databoot= rmultinomial(rep(NN, 10000),p)
m00boot =theta* (databoot[,2]*databoot[,3])/databoot[,1]
nboot = databoot[,1:3]
Nboot = m00boot + nboot[,1]+ nboot[,2]+ nboot[,3]
quantile(Nboot, c(0.025, 0.5, 0.975), type = 1)
sd = function(x) sqrt(var(x))
sd(Nboot)
```

3

The effects of imperfect linkage and erroneous captures

Susanna C Gerritse

Statistics Netherlands

Bart F.M. Bakker

Statistics Netherlands/VU University

Daan B Zult

Statistics Netherlands

Peter G.M. van der Heijden

Utrecht University/University of Southampton

CONTENTS

3.1 Introduction

Capture-recapture methods are commonly used to estimate the size of hard-to-reach populations Fienberg (1972); Bishop et al. (1975); Cormack (1989); International Working Group for Disease Monitoring and Forecasting (1995). When linking the individuals from two or more registers we can use these methods to estimate the portion of the population that was missed by all registers.

Capture-recapture estimation relies on five assumptions: 1) For the two-register case, the registers are assumed to be independent in the sense that the inclusion probability of one register is independent of the inclusion probability of the other register. For three registers this assumption is relaxed and it is only assumed that the three-factor interaction is zero, such that dependence between pairs of two registers may occur; 2) The registers are perfectly linked: when one unit is captured in two (or more) registers, perfect linkage assumes that we correctly identify all of these units as recaptures; 3) The population is closed: registers with continuous recording such as a population register are closed when one point in time is chosen. For incidence registers it is recommended to take a small sampling period to limit a possible violation of the closed population assumption; 4) All individuals in the registers belong to the population, i.e. there are no erroneous captures; 5) Assumptions related to homogeneity of inclusion probabilities (Van der Heijden et al., 2012). Heterogeneity occurs when one register has heterogeneous inclusion probabilities, for example when the probability to include men is higher than the probability to include women. If there is one source of heterogeneity, the estimate is unbiased when at least for one of the two registers the inclusion probabilities are homogeneous (Chao et al., 2001; Zwane and Van der Heijden, 2007). If there is a source of heterogeneity in each of two registers, the estimates are unbiased if the inclusion probabilities of the two sources of heterogeneity are statistically independent (Gerritse et al., 2015b; Seber, 1982, p. 86).

S. C. Gerritse, B. F. M. Bakker, D. Zult, and P. G. M. van der Heijden. The effects of imperfect linkage and erroneous captures on the population size estimator. Submitted, 2016. Acknowledgement of author contributions: theoretical development by SG. BB and DB contributed to theoretical development. Document written by SG (main document) and DZ (part on simulation study), edited by PvdH and BB. Analyses carried out by SG (main document) and DZ (part on simulation study).

In capture-recapture methodology it is not possible to verify from the data whether the assumptions are met, let alone to verify from the data to what extent possible violations occur. Some research has been conducted to investigate the effect of violations of the assumptions of capture-recapture analysis on the population size estimate for two registers. For the independence assumption, it has been shown that violating the independence assumption in the two-register case can lead to biased results, but the bias is not necessarily large. This result is explained via the relative size of the implied coverage of the register. Assume that register 1 has a larger coverage of the population than register 2, such that some of the units from register 2 lead to an increased coverage of the population. Then implied coverage is the number of units included in the first register also being included in the second register, relative to the size of the second register. High implied coverage results from a high number of units being included in the second register also being included in the first register. When the data has a high implied coverage the capture-recapture estimation is relatively robust to a violation of independence between the registers (Brown et al., 1999; Boden, 2014; Gerritse et al., 2015b).

Boden (2014) used sensitivity analyses to test the effect of violating the assumptions of independence, perfect linkage and heterogeneity. The sensitivity analysis conducted on a violation of independence between the sources showed that dependence between the registers had a substantial effect on the population size estimate. However, the results of violating perfect linkage and heterogeneity showed only a small bias. The small effect of linkage error on the capture-recapture estimation may be explained by the small number of linkage errors in the paper, although the author concludes that in his case even a moderate amount of linkage errors would have a minor impact on the population size estimate.

In this paper we investigate the effect of linkage error and erroneous captures on the population size estimation, assuming all other assumptions have been met. For this purpose, we use administrative data from Statistics Netherlands: the Population Register (PR), an Employment Register (ER) and a Crime Suspects Register (CSR), all from 2010. The three registers used have been linked for the most part via deterministic linkage, made possible by the use of a unique Personal Identification Number (PIN) that all PR registered individuals in the Netherlands have. When there are no administrative errors, this PIN can identify every individual correctly. However, these registers may contain administrative errors such that deterministic linkage may not identify all links. To be able to identify these links we also used probabilistic linkage. Probabilistic linkage estimates for each possible pair of individuals in two registers their probability of a correct link (Fellegi and Sunter, 1969).

Those individuals in the ER and CSR that did not already link to the PR were linked probabilistically. The advantage of this method is that it allows for small errors in the identifying variables.

By using both deterministic and probabilistic record linkage, after linkage it was found that 37 percent of the individuals in the CSR that did not link to the PR and ER had missing information on one of the linkage variables, and could not be linked. It is possible that these individuals do not belong to the population and it is impossible to know whether they should have been linked or not. Thus there is the possibility of two assumptions being violated: perfect linkage and no erroneous captures. There is also a possibility that part of the data of the CSR that have missing data are duplicate units, thus one person who is registered twice but can not be identified as the same person due to missing information. We assume in this paper that duplicate units do not occur.

Since the number of captured and recaptured units are used to estimate the number of units missed by the registers for population size estimation, the process of record linkage is an important aspect of capture-recapture methodology when using administrative data. Errors in record linkage will result in violation of the perfect linkage assumption. There are two types of linkage errors. For the sake of simplicity we exemplify the possible errors in linkage for two registers only, register 1 and register 2, and two unique units, X and Y. One error in linkage occurs when unit X in register 1 is falsely linked to unit Y in register 2. This type of error is also known as a mislink or a false positive. A second error in linkage occurs when unit X in register 1 is falsely not linked with unit X in register 2. This type of error is known as a missed link or a false negative.

If there are no covariates involved, we are in a relatively simple situation where one false positive can be compensated by one false negative, and thus there will be no effect on the population size estimation. Thus linkage errors are the number of false positives minus the number of false negatives, or in other words the number of mislinks minus the number of missed links. Then linkage errors are seen as a balance of two possible errors. Throughout this manuscript we will most commonly refer to errors in record linkage as linkage error. Erroneous captures occur when units that do not belong to the population are in the data.

The three registers from Statistics Netherlands together contain data on individuals residing in the Netherlands. In this manuscript we study two nationality groups from the three registers. The first nationality group contains data from all three registers on one nationality only: Polish. In the EU, individuals with EU nationalities are free to move and work within the EU. Hence, Polish individuals are free to live in the Netherlands and a high number of the labor migrants in the Netherlands in 2010 were from Poland. The second nationality group contains data from all three registers on individuals with an Iraqi, Iranian and Afghan nationality. Individuals from these three countries need a visa and working permit to enter the Netherlands and are undocumented immigrants when residing in the Netherlands without either of these. These nationality groups differ substantially in their implied coverage: the Polish have a low implied coverage, whereas the Iraqi, Iranian and Afghan have a high implied coverage.

We continue as follows. In section 2, a sensitivity analysis is conducted on these two nationality groups for the simplest form of capture-recapture: using only two registers. Section 2.1 will discuss the effect of linkage error on the population size estimate and section 2.2 will discuss the effect of erroneous captures on the population size estimate. In section 3 we extend the sensitivity analysis to the multiple-register case, where three registers are used to exemplify one form of a multiple-register case. In section 3.1 a simulation study is carried out on the effect of linkage errors on the population size estimate. A simulation study was chosen to investigate which error percentage in any of the three registers will result in the highest bias to the population size estimate. Using the knowledge gathered in section 3.1, a sensitivity analysis will be conducted in section 3.2 on the effect of linkage errors on the population size estimate. In section 3.3 the effect of erroneous captures on the population size estimate is established via another sensitivity analysis. Section 4 will give a discussion of the results and we conclude this manuscript in section 5.

3.2 Capture-recapture for two registers

The simplest population size estimation model makes use of two registers: 1 and 2. Let variables A and B respectively denote inclusion in registers 1 and 2. Let the levels of A be indexed by i ($i = 0$ (No), 1 (Yes)) where $i = 0$ stands for "not included in register 1", and $i = 1$, stands for "included in register 1". Similarly, let the levels of B be indexed by j $(j = 0(No), 1(Yes))$. Expected values are denoted by m_{ij} and fitted values are denoted by $\hat{m}_{ij}$. Observed values are denoted by n_{ij}, with $n_{00} = 0$ by design.

The first assumption of population size estimation is that the probability of being in the first register is independent of the probability of being in the second register. Under independence the saturated loglinear model for the counts n_{01}, n_{10} and n_{11} is:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B. \tag{3.1}$$

where we used the identifying restrictions $\lambda_0^A = \lambda_0^B = 0$. Two ways to derive the maximum likelihood estimate of the missed part of the population, are first, by using Poisson loglinear modeling such that $\hat{m}_{00} = \exp(\hat{\lambda})$ and second, by using the property that the odds ratio under independence is 1, i.e., $m_{00}m_{11}/m_{10}m_{01} = 1$ so that:

$$\hat{m}_{00} = \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10}n_{01}}{n_{11}}. \tag{3.2}$$

Poisson loglinear modeling is more flexible in more complicated loglinear models using covariates. The odds ratio provides a simple trick to explore capture-recapture methodology, but becomes increasingly difficult to use as the number of registers and covariates increases (compare, Gerritse et al., 2015b).

3.2.1 Linkage errors in two registers, theoretical example

An important assumption of capture-recapture methodology is perfect linkage. The assumption is met when all units in register 1 that are also in register 2 are correctly linked to their counterparts in register 2 and when all units in register 2 that are also in register 1 are correctly linked to their counterparts in register 1. We are interested in what happens if the two registers have not been linked perfectly when capture-recapture analysis is done assuming the assumptions to be met.

When linking two registers the contingency table with the expected values m_{ij} are shown in the left-hand side of Table 3.1, where m_{00} will be estimated via a maximum likelihood estimate (3.2). The population size estimate under assumed perfect linkage is $\hat{N} = \hat{m}_{11} + \hat{m}_{01} + \hat{m}_{10} + \hat{m}_{00} = n_{11} + n_{01} + n_{10} + \hat{m}_{00}$.

Assume we have linkage errors of size b, where b is the number of false positive links minus the number of false negative links. In short we get: linkage error = (b =) number of false positive links - number of false negative links. Then b is negative when the number of false negative links outbalances the number of false positive links, and b is positive when the number of false positive links outbalances the number of false negative links.

Under perfect linkage we assume that $b = 0$ for expected values m_{ij} and observed values n_{ij}. Under linkage error $b \neq 0$ we denote expected

TABLE 3.1
Expected values of being present in register 1 and register 2 under perfect linkage on the right side of the table and the expected values of being present in register 1 and register 2 under linkage error on the left side.

	A				A		
B	1 (yes)	0 (no)	Total	B	1 (yes)	0 (no)	Total
1 (yes)	m_{11}	m_{10}	m_{1+}	1 (yes)	$\tilde{m}_{11}$	$\tilde{m}_{10}$	$\tilde{m}_{1+}$
0 (no)	m_{01}	m_{00}	m_{0+}	0 (no)	$\tilde{m}_{01}$	$\tilde{m}_{00}$	$\tilde{m}_{0+}$
Total	m_{+1}	m_{+0}	m_{++}	Total	$\tilde{m}_{+1}$	$\tilde{m}_{+0}$	$\tilde{m}_{++}$

values by $\tilde{m}_{ij}$, and observed values are denoted $\tilde{n}_{ij}$. Then $\tilde{m}_{11} = m_{11}+b$, $\tilde{m}_{10} = m_{10} - b$ and $\tilde{m}_{01} = m_{01} - b$, and $\tilde{n}_{11} = n_{11} + b$, $\tilde{n}_{10} = n_{10} - b$ and $\tilde{n}_{01} = n_{01} - b$. The contingency table with the expected values $\tilde{m}_{ij}$ is shown in the right-hand side of Table 3.1.

Unfortunately, we can not verify from the data to what extent perfect linkage has been violated. We can however use chosen values of b to investigate the effect of linkage error on the population size estimate in a sensitivity analysis. To choose values of b we define linkage error rate β for m_{01}. For this specific example we have chosen linkage error rate β on m_{01} because m_{01} is the number of added cases of register 2 relative to register 1, such that β is specified based on the implied coverage of register 1 given register 2. Then,

$$\beta = \frac{\tilde{n}_{01}}{n_{01}} \tag{3.3}$$

where $\beta = 1$ denotes perfect linkage. Linkage error rate β enables us to simulate linkage error, where $\tilde{n}_{01} = n_{01} * \beta$. In creating such a linkage error rate β we can denote linkage error in percentages, and by defining linkage error in percentages we can better compare the effect of β on the population size estimate between the two nationality groups. Then,

$$\frac{\hat{\tilde{m}}_{10}\hat{\tilde{m}}_{01}}{\hat{\tilde{m}}_{11}} = \frac{\tilde{n}_{10}\tilde{n}_{01}}{\tilde{n}_{11}} = \frac{(n_{10} - b)(n_{01} - b)}{n_{11} + b} = \hat{m}_{00(\beta)}, \tag{3.4}$$

where $\hat{m}_{00(\beta)}$ is the size of the individuals missed by the two registers. The population size estimate under linkage error is $\hat{N}_{\beta} = \hat{m}_{00(\beta)} + (n_{11} + b) + (n_{10} - b) + (n_{01} - b)$.

It can be shown that if b is positive, and thus the number of false positive links outbalances the number of false negative links, $\hat{m}_{00(\beta)}$ will be smaller than $\hat{m}_{00}$ and $\hat{m}_{00}$ is an overestimation of $\hat{m}_{00(\beta)}$. If b is negative, and thus the number of false negative links outbalances the number of false positive links, $\hat{m}_{00(\beta)}$ will be larger than $\hat{m}_{00}$ and $\hat{m}_{00}$ is an underestimation of $\hat{m}_{00(\beta)}$. Then, when we estimate $\hat{m}_{00}$ while we have a linkage error of size β, estimate $\hat{m}_{00}$ will be biased.

Note that n_{1+} and n_{+1} are fixed values, because these are the total number of individuals in register 1 and register 2. Under linkage error the number of individuals for register 1 and register 2 does not change. However, the sizes of the expected values $\tilde{m}_{ij}$ do change.

3.2.2 Implied coverage

Under register 1 and 2, the maximum likelihood estimate of the missed portion of the population can be estimated via equation(3.2). Under 3.2 we can estimate conditional probabilities:

$$\hat{p}(0|1) = \frac{n_{01}}{n_{+1}} \text{ and, } \hat{p}(1|1) = \frac{n_{11}}{n_{+1}}, \tag{3.5}$$

where $\hat{p}(0|1)$ is the estimated probability of n_{01}, given n_{+1}: the conditional, estimated probability of only being registered in register 2, given all registered cases in register 2, including the overlap with register 1. Similarly, $\hat{p}(1|1)$ is the estimated probability of n_{11}, given n_{+1}: the conditional, estimated probability of being in the overlap between register 1 and register 2, given all registered cases in register 2. Thus $\hat{p}(0|1)$ is the estimated probability of new cases from register 2, compared to those cases already registered in register 1, and $\hat{p}(1|1)$ is the estimated probability of already known cases from register 1, compared to all the cases from register 2. These two probabilities together add up to 1. Given these probabilities we can rewrite equation (3.2)

$$\hat{m}_{00} = \frac{n_{10} * \hat{p}(0|1)}{\hat{p}(1|1)}, \tag{3.6}$$

where the number of observations uniquely in register 1, multiplied with the estimated odds of a new observation found in register 2. The larger the estimated odds, the larger $\hat{m}_{00}$ will be. It can be seen from equation (3.6) that the estimated number of individuals missed by the two registers is a result of the estimated probability of new cases added by register 2, compared to the number of cases already known in register 1.

When the estimated probability of new cases $\hat{p}(0|1)$ is relatively small compared to the estimated probability of already known cases $\hat{p}(1|1)$, the effect of the added new cases of register 2 on $\hat{m}_{00}$ will be small and the population size estimator is robust. Then the coverage of the population by register 1, implied by register 2 is high, which means that register 1 already captures a high number of the individuals in the population compared to register 2.

However, when the estimated probability of new cases $\hat{p}(0|1)$ is relatively large compared to the estimated probability of known cases $\hat{p}(1|1)$, the effect of the added new cases of register 2 on $\hat{m}_{00}$ will be large as well and the population size estimator is not robust. Then the coverage of the population by register 1, implied by register 2 is low, such that register 2 captures a relatively larger number of unique cases compared to register 1. The coverage of register 1 is implied by register 2 because these two registers are our reference point, and the true coverage is unknown. In this paper we will denote the coverage of register 1 implied by register 2 as implied coverage.

3.2.3 Linkage error for two registers, real data examples

To illustrate, we use the data on individuals with a Polish nationality residing in the Netherlands, and data of individuals with anAfghan, Iraqi and Iranian nationality residing in the Netherlands For this section, where we investigate the effect of linkage error on the simplest capture-recapture case with two registers only, we use two of the three registers introduced in the introduction: the Dutch Population Register (PR) and a Crime Suspect Register (CSR). The data are shown in Table 3.2.

TABLE 3.2
Left are the values of the Afghan, Iraqi and Iranian individuals and on the right are values of the individuals of Polish nationality, estimated values are in italics.

	CSR		
PR	Yes	No	Total
Yes	1,356	58,891	60,247
No	320	*13,898*	*14,218*
Total	1,675	*72,789*	*74,465*

	CSR		
PR	Yes	No	Total
Yes	444	42,109	42,553
No	1,659	*157,340*	*158,999*
Total	2,103	*199,449*	*201,552*

We illustrate the use of β and b by using data of individuals with an Afghan, Iraqi and Iranian nationality. Assuming perfect linkage under $\beta = 1$, we get a maximum likelihood estimate of the missed part of the population by $58,891 * 320/1,356 = 13,898$. This gives a total of $\hat{N} = 1,356 + 58,891 + 320 + 13,898 = 74,465$ individuals with an Afghan, Iraqi and Iranian nationality are residing in the Netherlands.

However, when we introduce a linkage error of size $\beta = 0.9$, then $\tilde{n}_{01} = n_{01} * 0.9 = 320 * 0.9 = 288$, and $b = 320 - 288 = 32$, such that $\tilde{n}_{10} = 58,891 - 32 = 58,859$ and $\tilde{n}_{11} = 1,356 + 32 = 1,388$, which can be seen on the right-hand side of Table 3.3. We estimate that 12,213 individuals are missed by all three registers but do belong to the population. Theestimate of the missed portion of the population allows us to get estimate $\hat{N}_\beta = 1,388 + 58,859 + 288 + 12,213 = 72,748$ individuals with an Afghan, Iraqi and Iranian individuals actually residing in the Netherlands. Thus the estimate of 74,465 is a result of capture-recapture analysis when assuming perfect linkage, and will be an overestimation if we have linkage error and have not adjusted for this.

TABLE 3.3
The left table are the observed values for the CSR for the Afghan, Iraqi and Iranian people residing in the Netherlands, the right table are the observed values under error linkage $\beta = 0.9$. Estimated values are in brackets.

	CSR		
PR	Yes	No	Total
Yes	1,356	58,891	60,247
No	320	*13,898*	*14,218*
Total	1,675	*72,789*	*74,464*

	CSR		
PR	Yes	No	Total
Yes	1,388	58,859	60,247
No	288	*12,213*	*12,501*
Total	1,676	*71,072*	*72,748*

Parameter β enables us to conduct a sensitivity analysis on linkage errors. The results can be seen in Figure 1 and details can be found in the Appendix. When a linkage error range of $\beta = 0.5$ to $\beta = 1.5$ is introduced on the data of the Afghan, Iraqi and Iranian individuals, the estimate of the missed portion of the population does not differ greatly from the estimate under perfect linkage. If we introduce a linkage error of size $\beta = 0.5$ (and also for $\beta = 1.5$) and estimate assuming perfect linkage without adjusting for the linkage error, we see a bias of only 12 percent. Thus there is only a 12 percent difference between the actual population size estimate $\hat{N}_\beta$ and the population size estimate under the observed values $\hat{N}$ where perfect linkage is assumed. Thus for the Afghan, Iraqi and Iranian individuals the population size estimator is relatively robust.

To compare we also conducted a sensitivity analysis on the effect of linkage error on the individuals with a Polish nationality. We introduced a linkage error range of $\beta = 0.5$ to $\beta = 1.5$, where the upper range of β is lower because of the low cell count of 444 in cell (1,1). As can be seen from the figure, the population size estimator is not robust to linkage error. Introducing a linkage error of $\beta = 0.5$, we overestimate the population size estimate by 187 percent, compared to when we would estimate under $\beta = 1$. While the population size estimate under linkage errors is quite stable for the individuals with an Afghan, Iraqi and Iranian nationality, for the individuals with a Polish nationality this is not the case.

Thus for the Afghan, Iraqi and Iranian individuals the population

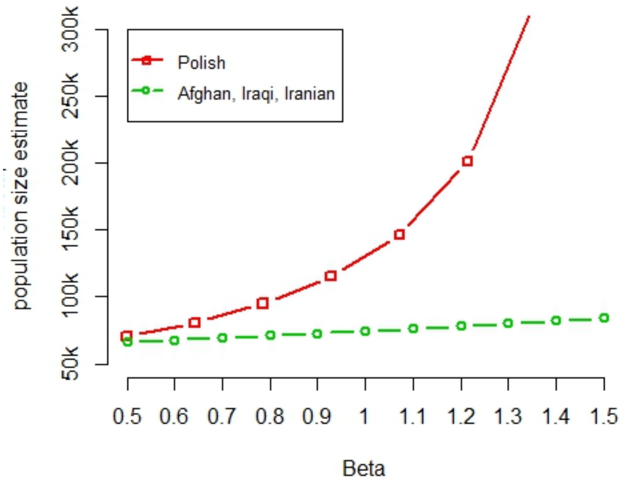

FIGURE 3.1
Population size estimate for the two-register capture-recapture sensitivity analysis for both nationalities under a β ranging from 0.5 to 1.5. Under $\beta = 1$ perfect linkage is assumed and for $\beta \neq 1$ linkage errors are assumed.

size estimator is relatively robust under linkage errors, but for the Polish individuals the population size estimator is not robust. The reason is that for the former, there is a larger implied coverage of the PR given the CSR than for the latter. For the Afghan, Iraqi and Iranian individuals $\hat{p}(0|1) = 320/1,675 = 0.19$ and $\hat{p}(1|1) = 1,356/1676 = 0.81$, which means that the CSR does not add many new cases to the PR, and the estimated conditional coverage implied by the PR given the CSR is high. For the Polish individuals $\hat{p}(0|1) = 1,659/2,103 = 0.79$ and $\hat{p}(1|1) = 444/2,103 = 0.21$, which means that the CSR actually adds many new cases to the PR, and the estimated conditional coverage implied by the PR given the CSR is low. For the Afghan, Iraqi and Iranian individuals, due to the high implied coverage, the population size estimator is robust. Whereas for the Polish individuals, due to the low implied coverage, the population size estimator is not robust at all.

3.2.4 Erroneous captures for two registers, theoretical example

Another important assumption of capture-recapture methodology is that all the individuals in the observed data belong to the population, and thus the data contain no erroneous captures. It is not always possible to assess from the data which units actually belong to the population and which do not. Therefore we aim to investigate the effect on the population size estimate when erroneous captures are present.

Again m_{ij} are the expected values of the observed values n_{ij}. For m_{ij} and n_{ij} we assume no erroneous captures. If erroneous captures are introduced, expected values are denoted by $\bar{m}_{ij}$, and observed values under linkage error are $\bar{n}_{ij}$. We can define an erroneous capture rate γ for n_{01}, where $\gamma = \bar{n}_{01}/n_{01}$, such that $\bar{n}_{01} = n_{01} * \gamma$. Erroneous captures are units in the data that should not have been observed, and therefore $\bar{n}_{01}$ will always be smaller than n_{01}, and $0 \leq \gamma \leq 1$. The erroneous capture rate γ has been defined on n_{01} because that is the number of added cases by register 2, relative to register 1. We find

$$\hat{m}_{00(\gamma)} = \frac{n_{10}(n_{01*}\gamma)}{n_{11}} = \frac{n_{10}\bar{n}_{01}}{n_{11}} = \gamma \hat{m}_{00}, \qquad (3.7)$$

where $\hat{m}_{00}$ is the estimate when there are no erroneous captures defined in (3.2). This enables us to choose values for γ and set up sensitivity analyses to the effect of erroneous captures on the population size estimator.

3.2.5 Erroneous captures for two registers, real data example

For the sensitivity analyses to study the effect of erroneous captures on the population size estimator, we again use the data from Table 3.2. We chose the relatively extreme range $\gamma = 0.9$ to $\gamma = 0.1$. Such extreme ranges may not be very realistic, but they do give us a complete picture of the effect of γ. Table 3.4 shows the expected values when there are 10 percent erroneous captures in n_{01} and hence $\gamma = 0.9$. Then $\bar{n}_{01} = n_{01} * \gamma = n_{01} * 0.9 = 320 * 0.9 = 288$.

TABLE 3.4
The values for the Afghan, Iraqi and Iranian nationals residing in the Netherlands in 2010, adjusted for $\gamma = 0.9$ erroneous capture in the CSR. Values in italics are estimated values

	CSR		
PR	Yes	No	Total
Yes	1,356	58,891	60,247
No	288	*12,508*	*12,796*
Total	1,644	*71,399*	*73,043*

It follows that, $\bar{m}_{00(\gamma)} = 12,508$, which is 1,390 less individuals estimated to be missed by all three registers than $\hat{m}_{00} = 13,898$. Figure 2 shows the sensitivity analysis for both nationality groups. For the individuals with an Afghan, Iraqi and Iranian nationality, erroneous captures have only a small an effect on the population size estimator and the estimator is relatively robust. Figure 2 also shows the sensitivity analysis for the Polish individuals. For these individuals, erroneous captures have a large effect on the population size estimator and the estimator is not robust.

Here again the population size estimator is more robust against violation of an assumption for the data of the Afghan, Iraqi and Iranian individuals than for the Polish individuals. The effect of erroneous captures is smaller than the effect of linkage error, because it only affects one expected value, whereas linkage errors affect all three cells, compare (3.4) and (3.7).

Here again the relative size of the implied coverage of the PR given the ER is the explanation why one nationality group results in less bias in the population size estimate after violated assumptions than the other nationality group. Given that the erroneous captures are deleted from n_{01}, the adjusted size of n_{01} will influence the implied coverage of reg-

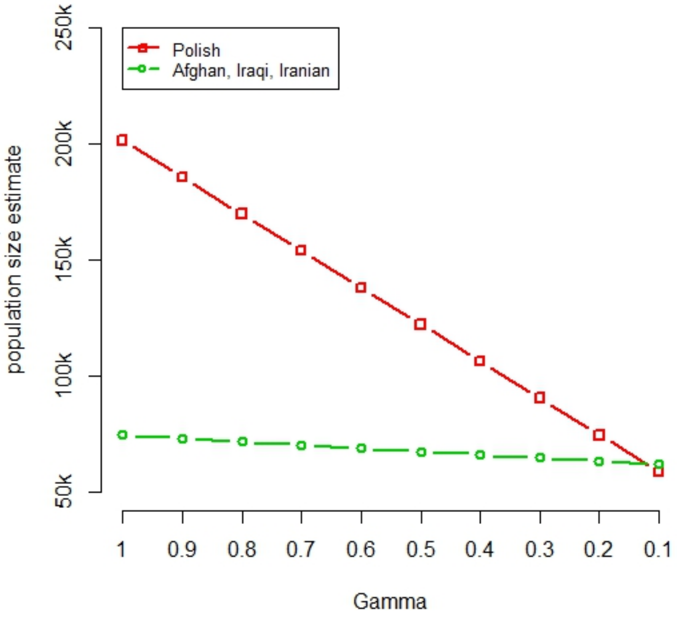

FIGURE 3.2
Population size estimate for the two-register capture-recapture sensitivity analysis for both nationality groups under a Gamma ranging from 1 to 0.1. Under $\gamma = 1$ we assume no erroneous captures, and when $\gamma<1$ the data contain erroneous captures.

ister 1 given register 2. For the Afghan, Iraqi and Iranian individuals $\hat{p}(0|1) = 320/1,675 = 0.19$, but when $\gamma = 0.9$, $n_{01} = 288$ instead of 320 and $\hat{p}(0|1) = 288/1,675 = 0.17$. Thus when eliminating the erroneous captures from n_{01}, the implied coverage of the PR given the CSR increases.

3.3 Capture-recapture for three registers.

In section 2 we discussed the impact of linkage errors and erroneous captures in the simplest form of capture-recapture for two registers. However, multiple registers are often available. Additionally in using a third register, the strict independence assumption can be replaced by a less strict independence assumption where independence is assumed on the third register. Thus the use of a third register is not only often available to the researcher, but also advisable. When linkage errors or erroneous captures are present in three registers, the effect of these violations may become more complex. We now investigate for three registers what the effect is of violations of perfect linkage and no erroneous captures on the population size estimator.

Assume we have three registers, 1, 2 and 3. Let variables A, B and C respectively denote inclusion in registers 1, 2 and 3. Let the levels of A be indexed by i (i = 0,1) where $i = 0$ stands for "not included in register 1", and $i = 1$, stands for "included in register 1". Similarly, let the levels of B be indexed by j ($j = 0, 1$), and let the levels of C be indexed by k ($k = 0, 1$). Table 3.5 shows the expected values denoted by m_{ijk}. Observed values are denoted by n_{ijk} with $n_{000} = 0$.

TABLE 3.5
The table of expected counts for thee registers

		C	
A	B	1 (Yes)	0 (No)
1 (Yes)	1 (Yes)	m_{111}	m_{110}
	0 (No)	m_{101}	m_{100}
0(No)	1 (Yes)	m_{011}	m_{010}
	0 (No)	m_{001}	m_{000}

For three variables the saturated loglinear model is denoted by:

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}, \qquad (3.8)$$

with identifying restrictions that a parameter equals zero when i, j or $k = 0$. The model assumption is that the three-factor interaction parameter $\lambda_{ijk}^{ABC} = 0$. Model $[AB][BC][AC]$ is the saturated model, as the number of observed parameters equals the number of parameters to be estimated. This model assumes that the odds ratio between A and B is the same for $k = 0$ and $k = 1$, i.e.,

$$\frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}}. \tag{3.9}$$

An estimate for $\hat{m}_{000}$ is easily derived from (3.9).

$$\frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}} = \hat{m}_{000}. \tag{3.10}$$

When linkage error is present, the same rules apply as for two registers. Here again we can define a linkage rate β, where we investigate linkage error only in one register. As an example, linkage error rate β is only investigated in register 3:

$$\beta = \frac{\tilde{n}_{001}}{n_{001}}. \tag{3.11}$$

Under these conditions, we can again conduct a sensitivity analysis on data where the three registers are linked. To assess the effect oflinkage error on the population size estimator we first conduct a simulation study to investigate in which register linkage error will have more effect. This will help determine in which register linkage errors will have more effect on the population size estimate, so we can then determine one β. For the sensitivity analysis on real data we use this β to assess the effect of linkage errors on population size estimation for the three-register case.

3.3.1 Linkage error for three registers, simulation study

We aim to investigate the effect of linkage error on the population size estimator when three registers are used for capture-recapture analysis. However the combination of possible linkage errors for three registers becomes numerous and complex. For that reason we conduct a Monte Carlo (MC) simulation to assess the effect of different linkage errors on the population size estimator for three registers.

Table 3.6 shows the parameters used to generate simulated data. The size of the registers are the actual sizes in the PR, CSR and the ER for individuals without a Dutch nationality. Note that this encompasses but is not limited to the two nationality groups used before. The error percentages are chosen as best resembling the probable error percentages in these registers. Additionally the three chosen true population sizes are

TABLE 3.6
MC simulation parameters and their settings

Parameters MC simulation	Value
Number of MC iterations per scenario	25
Size of register 1	617,332
Size of register 2	374,803
Size of register 3	12,419
True population	0.7, 1 and 1.3 million
Error percentage register 1	0, 0.5, 2, 10
Error percentage register 2	0, 0.5, 2, 10
Error percentage register 3	0, 10, 40

chosen as the most probable true population sizes. The number of MC iterations is set at 25.

First a data set is generated, where we set the size of the true population at either 700 thousand, 1 million or 1,3 million individuals. These three population sizes have been chosen because they are deemed the most probable to be the number of individuals without a Dutch nationality residing in the Netherlands. Then register 1 is generated by randomly sampling 617,332 units from the true population, register 2 is generated by randomly sampling 374,803 units from the simulated data set and register 3 is generated by randomly sampling 12,419 units from the true population. Note that by randomly sampling the registers from the simulated data we are operating under loglinear model [A][B][C].

To assess the effect of linkage error on the population size estimator, we destroy links by changing the linkage key according the error percentages specified in Table 3.6. By renaming linked units, the number of units in register 1, register 2 and register 3 remain constant, but records that should have been linked no longer do. In doing so we also introduce interactions between the registers. These are accounted for by using model [AB][AC][BC] in the capture-recapture analysis.

3.3.2 MC simulation results

Together the parameter settings from Table 3.6 lead to 144 (3 x 4 x 4 x 3) different scenarios. For each scenario 25 MC iterations are performed, which implies that for each scenario, 25 times 3 registers are generated with corresponding error shares and equal capture-recapture population estimates. A relevant selection is taken from all computations and is shown in Table 3.7.

TABLE 3.7
A selection of the MC simulation results

	Scenario 0	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Scenario 9
Population size	1 mln	1 mln	1 mln	1mln	0.7 mln	1mln	1.3 mln	1 mln	1 mln	1 mln
Error % register 1	0%	10%	0%	0%	0.5%	0.5%	0.5%	0.5%	10%	10%
Error % register 2	0%	0%	10%	0%	2%	2%	2%	2%	2%	10%
Error % register 3	0%	0%	0%	10%	40%	40%	40%	10%	10%	10%
Unobserved population	236,361	296,336	236,196	236,247	37,740	236,228	481,163	236,249	236,284	236,289
Estimated unobserved population	234,401	280,108	307,636	346,153	590,485	949,575	1,414,345	363,354	427,606	535,139
Estimate of total population	998,010	1,043,772	1,071,450	1,109,906	1,252,852	1,713,425	2,233,240	1,127,143	1,191,266	1,301,293
Linking errors between 1 and 2	0%	10%	10%	0%	2.5%	2.5%	2.5%	2.5%	12%	19%
Linking errors between 1 and 3	0%	10%	0%	10%	40%	40%	40%	10%	19%	19%
Linking errors between 2 and 3	0%	0%	10%	10%	41%	41%	41%	12%	12%	19%
Overestimation % of population	-1%	4%	7%	11%	79%	71%	72%	13%	19%	30%

Row (1) from Table 3.7 is the true population size. Row (2) through (4) contain the error percentages that are introduced in the corresponding registers. Row (5) contains the actual size of the population missed by all three registers. Row (6) contains the estimated size of the population missed by all three registers. Under perfect linkage the numbers in (5) and (6) should be similar. Rows (7) through (10) give descriptive statistics, i.e. an estimation of the total population and the percentages of linking errors between the three registers as a result of the error percentages in the registers. Finally, row (11) contains the difference (in percentages) between the estimated population size estimate and real population size estimate.

Scenario 0 shows the capture-recapture estimation when no linkage errors are introduced into the simulation study. As can be seen, scenario 0 estimates the missed part of the population with a slight but negligible underestimation. The next three scenarios differ only with respect to the register in which errors occur. However, despite their similarity, row (11) shows that there is a difference between the quality of their estimates. There is a clear relationship between the bias in population size estimate relative to the actual population size and the register in which the error manifests itself. Interestingly, the bias in population size estimation increases when the errors occur in a smaller register. A ten percent linkage error in a large register biases the population size estimate less than a ten percent linkage error in a small register, even though the absolute number of linkage errors in a large register is larger than the absolute number of linkage errors in the smaller register.

The next three scenarios (4 through 6) differ only with respect to their real population size. In relative terms the difference between the estimate of the population size from capture-recapture analysis and the actual pre-defined population size increases when the population gets smaller; in absolute terms the relationship is reversed (with larger population sizes come larger errors). In the last three columns there are three scenarios that have linkage error percentages which are potentially similar to the Dutch situation, with a varying real population.

3.3.3 Linkage error in three registers, real data example.

The simulation study has shown that the largest bias in population size estimation results from errors in the smallest register, which in this example is the CSR. Thus for our sensitivity analysis we focus on the CSR, and we use β for our sensitivity analysis on actual data. We again use data on the Afghan, Iraqi and Iranian individuals residing in the Netherlands, which is shown on the left side of Table 3.8. Additionally, we again use the data on the Polish individuals residing in the Netherlands, shown on the right side of Table 3.8.

TABLE 3.8
The observed values for the Afghan, Iraqi and Iranian individuals on the left side of the table and the observed values for the Polish individuals on the right side of the Table by three registers.

		CSR	
PR	ER	Yes	No
Yes	Yes	309	13,862
	No	1,047	45,029
No	Yes	7	266
	No	313	0

		CSR	
PR	ER	Yes	No
Yes	Yes	215	24,832
	No	229	17,277
No	Yes	230	80,406
	No	1,429	0

We introduce the linkage error rate $\beta < 1$, where individuals in the CSR are to be linked to either the ER or the PR. Some of the CSR individuals who did not link to the PR will also partially have to be linked to the intersection of PR and ER. Then for the individuals in the CSR who will be linked to the PR, some will also be linked to the intersection of PR and ER. The proportion of linkage errors in the CSR to be linked to either the PR or the intersection of the PR and the ER will be the same as the distribution of the CSR individuals who are in the PR. 309 individuals with an Afghan, Iraqi and Iranian nationality are in the intersection of all three registers, compared to 1,647 in the PR and CSR, which is only 16 percent of all the individuals in both the PR and the CSR. The percentage linkage error of size β that should have linked to the PR will link for 16 percent to the intersection of the PR and the ER and for 84 percent to the PR alone.

As can be seen from Figure 4, the linkage error for the Afghan, Iraqi and Iranian individuals has again only a small effect on the population size estimation. However, for the Polish individuals linkage errors have a large effect on the population size estimation. Table 3.14 in the Appendix shows the estimates for the full sensitivity analysis. However, compared to the Polish individuals, the capture-recapture analysis for the Afghan, Iraqi and Iranian individuals is quite robust under the three registers.

The difference between the effect linkage errors have on the population size estimate between the nationalities again are the result of the implied coverage. In the sensitivity analysis we link individuals from the CSR to the PR and the ER. However, for the Afghan, Iraqi and Iranian individuals there are only 313 individuals in the CSR that did not link to the PR and ER, which is 1,429 cases for the Polish individuals. The coverage of the PR and ER is smaller for the Polish individuals com-

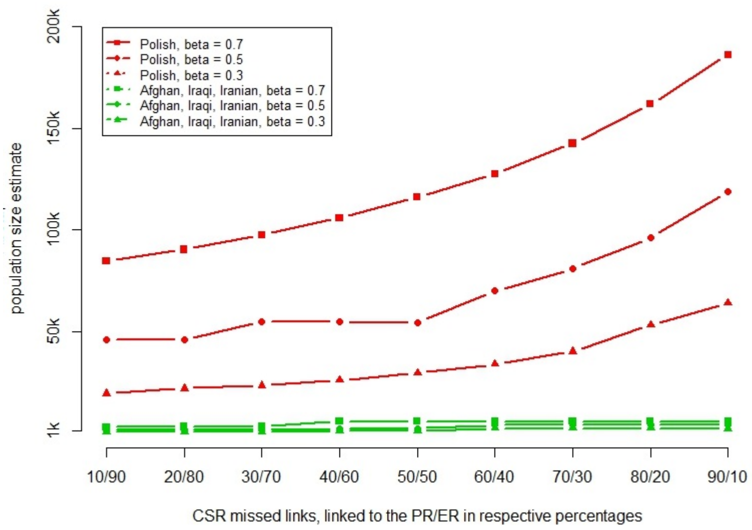

FIGURE 3.3
Population size estimate for both nationalities with linkage error rate $\beta = 0.7$, 0.5 or 0.3. The respective b has been distributed over the PR and ER according the percentages on the X-axis. For example, the first tick on the X axis, 10/90, means 10 percent of b were linked to the PR and 90 percent of b to the ER. Note that part of b linked to the PR, a part is distributed to the intersection of PR and ER.

pared to the Afghan, Iraqi and Iranian individuals, given 1,429 cases are new cases added by the CSR compared to only 313. Thus for the Polish individuals the population size estimate is less robust than for the Afghan, Iraqi and Iranian individuals.

3.3.4 Erroneous captures in three registers, real data example

When three registers are used and one register contains erroneous captures, a sensitivity analysis with γ can be carried out. We present here one example. We define an erroneous capture rate $\gamma = \bar{n}_{001}/n_{001}$. Erroneous capture rate γ was introduced for $\bar{n}_{001}$, because the CSR has the most administrative error. Thus the added cases of the CSR relative to the PR and ER may have the highest probability of erroneous captures. For the sensitivity analysis a range from $\gamma = 0.9$ to $\gamma = 0.1$ of erroneous captures are introduced to the observed values.

The bias resulting from erroneous captures in the capture-recapture analysis for both nationalities can be found in Figure 4. For erroneous captures in three registers, the population size estimate for the individuals with a Polish nationality is not robust to violate this assumption. When assuming that all the individuals belong to the population, the total number of Polish individuals in the Netherlands is 450,945. When, however, erroneous captures of $\gamma = 0.5$ are introduced, the population size is 286,953. Thus we overestimate the population size estimate by 65 percent when erroneous captures are present.

Again we find that when violating erroneous captures the population size estimate for the individuals with an Afghan, Iraqi and Iranian nationality is quite robust to erroneous captures. Under no erroneous captures the total number of Afghan, Iraqi and Iranian individuals in the Netherlands is 68,682. When, however, we operate under $\gamma = 0.5$ and we have 50 percent of erroneous captures in the CSR, the population size estimate is 64,889, and we overestimate by only six percent.

As was stated in the previous section, the implied coverage of the PR and ER is higher for the Afghan, Iraqi and Iranian individuals than for the Polish individuals. As such, the population size estimate under erroneous captures is more robust for the Afghan, Iraqi and Iranian individuals than for the Polish individuals.

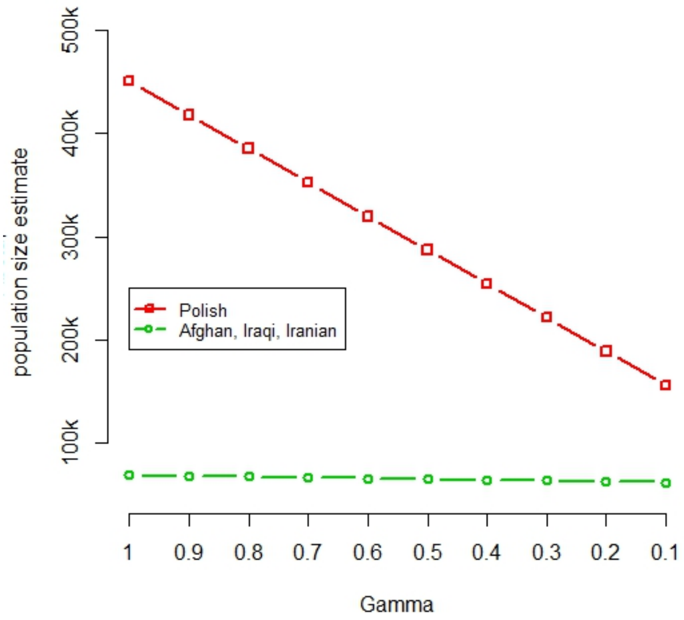

FIGURE 3.4
Population size estimate for both nationalities under a Gamma ranging from 1 to 0.1. Under $\gamma < 1$ no erroneous captures are assumed, and for $\gamma \neq 1$ erroneous captures are assumed.

3.4 Conclusion

In this manuscript we have compared two rather different nationality groups: data of Afghan, Iraqi and Iranian individuals residing in the Netherlands are compared to data of Polish individuals residing in the Netherlands to assess the effect of linkage errors and erroneous captures on the population size estimate. These two different nationality groups have been chosen because their implied coverage is different due to different legal requirements to reside in the Netherlands. This results in two rather different contingency tables, as can be seen from Table 3.2 and 3.8. Both nationality groups have a high number of individuals registered in the PR only. However, for the individuals registered in the CSR there is a large difference between the nationality groups on the implied coverage of the PR over the CSR.

Because individuals with a Polish nationality are free to move and work in the EU, less Polish individuals registered in the CSR are also in the intersection with PR. Thus the implied coverage of the PR relative to the CSR is low. For the individuals with an Afghan, Iraqi and Iranian nationality we see the opposite. Because these individuals need a working or residence permit to enter the Netherlands, more individuals registered in the CSR are also in the intersection with the PR. This is probably due to the fact that those CSR registered individuals who are not registered in the PR are illegally residing in the Netherlands, whereas legally the Polish individuals only registered in the CSR are not illegally residing in the Netherlands.

For the two register case we see that because the implied coverage of the PR over the CSR is different between the two nationality groups, the effect of linkage errors and erroneous captures is more dramatic for the Polish data than for the Afghan, Iraqi and Iranian data. The implied coverage of the PR given the CSR for the Afghan, Iraqi and Iranian individuals is already relatively high, such that the population size estimate is more robust to violation of the assumptions. However, for the Polish data the implied coverage of the PR given the CSR is rather small, such that the population size estimator is less robust to violations of the same percentage as for the Afghan, Iraqi and Iranian individuals.

Given the implied coverage has a substantial impact on the population size estimation when assumptions are violated, it is important that all units are linked correctly. Currently there are some developments in the theory and practice of capture-recapture methods that aims at linkage error-unbiased estimates. One of which is the research from Consiglio and Tuoto (2015) based on Ding and Fienberg (1994) for probabilistic linkage, where they propose to use the estimated number of false positives and false negatives. For our current data we were unable to use their method, because of the 37% of individuals in the CSR who are without background information. These cases create unrealistic false positive and false negative probabilities. Additionally, so far this method is only developed for two registers.

We have assumed in this paper that when assessing the effect of either linkage error or erroneous captures on the population size estimate, all other assumptions are met. This is not very realistic. However, this did allow us to assess the effect of violating the assumptions on the population size estimator. To introduce more than one violation per data set would make the presentation more complex.

In this manuscript we have chosen relatively random linkage error rates and erroneous capture rates. These were chosen to assess the effect of minor deviations in perfect linkage and no erroneous captures to a more extreme deviation. Given that we can not assess the extent of deviation of the assumptions in observed data, we use sensitivity analyses to assess the effect of relative to extreme deviations.

We have found that implied coverage is an important aspect of the effect that violations of assumptions may have on the population size estimator. When the implied coverage of register 1 given register 2 is large, the population size estimate is relatively robust to violations of assumptions. However, when the implied coverage of register 1 given register 2 is small, the population size estimate is not robust to violations of the assumptions. We advise researchers to assess the implied coverage of the data used, because this will have an effect on the population size estimate.

3.5 Appendix

3.5.1 Tables for section 3.2.3

Results for the sensitivity analysis for linkage error under error rate β in two registers for the Afghan, Iraqi and Iranian individuals can be found in Table 3.9. The first row shows the estimate of the missed portion of the population under a β from 0.5 to 1.5. The second row gives the population size estimate $\hat{N} = m_{11} + m_{01} + m_{10} + \hat{m}_{00}$. The third row gives a relative bias, the bias of the estimate under assumed perfect linkage $\hat{N}$ to the estimate adjusted for linkage error $\hat{N}(\beta)$, where $\hat{N}_{(\beta)} = \hat{m}_{00(\beta)} + m_{11+b} + m_{10-b} + m_{01-b}$.

TABLE 3.9
Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with an Afghan, Iraqi and Iranian nationality.

β	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
$\hat{m}_{00(\beta)}$	6,199	7,603	9,070	10,605	12,213	13,898	15,665	17,522	19,475	21,531	23,699
$\hat{N}_{(\beta)}$	66,606	68,042	69,541	71,208	72,748	74,465	76,264	78,153	80,138	82,226	84,426
$\hat{N}/\hat{N}_{(\beta)}$	1.12	1.09	1.07	1.05	1.02	1	0.98	0.95	0.93	0.90	0.88

Results for the sensitivity analysis for linkage error under error rate β in two registers for the Polish individuals can be found in Table 3.10. The first row shows the estimate of the missed portion of the population under a β from 0.5 to 1.2, given that it was impossible to take more linkage error. The second row gives the population size estimate $\hat{N}$. The third row gives a relative bias, the bias of the estimate under assumed perfect linkage $\hat{N}$ to the estimate adjusted for linkage error $\hat{N}(\beta)$.

TABLE 3.10
Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with a Polish nationality.

β	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$\hat{m}_{00(\beta)}$	26,894	37,218	51,285	71,441	102,657	157,340	277,525	754,465
$\hat{N}_{(\beta)}$	70,278	80,766	94,999	115,321	146,703	201,552	321,903	799,009
$\hat{N}/\hat{N}_{(\beta)}$	2.87	2.50	2.12	1.74	1.37	1	0.63	0.25

3.5.2 Tables for section 3.2.5

Table 3.11 shows the robustness analysis for the Afghan, Iraqi and Iranian people considering erroneous captures of size γ in the CSR register.

Row $\hat{m}_{00(\gamma)}$ shows the maximum likelihood estimate under erroneous captures of size γ. The second row is the total population size estimate $\hat{N}_{(\gamma)} = \hat{m}_{00(\gamma)} + m_{10} + m_{01\gamma} + m_{11} = \hat{N} + (\gamma - 1)(\hat{m}_{00} + m_{01})$. The third row is the relative bias of $\hat{N}_{(\gamma)}$ to $\hat{N}$.

TABLE 3.11
Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with an Afghan, Iraqi and Iranian nationality when adjusting for γ erroneous captures.

γ	1	0.9	0.8	0.7	0.6
$\hat{m}_{00(\gamma)}$	13,898	12,508	11,118	9,728	8,339
$\hat{N}_{(\gamma)}$	74,465	73,043	71,621	70,199	68,778
$\hat{N}/\hat{N}_{(\gamma)}$	1	1.02	1.04	1.06	1.08

γ	0.5	0.4	0.4	0.2	0.1
$\hat{m}_{00(\gamma)}$	6,949	5,559	4,169	2,780	1,390
$\hat{N}_{(\gamma)}$	67,356	65,934	64,512	63,091	61,669
$\hat{N}/\hat{N}_{(\gamma)}$	1.11	1.13	1.15	1.18	1.21

Table 3.11 shows the robustness analysis for the Polish people considering erroneous captures of size γ in the CSR register. Row $\hat{m}_{00(\gamma)}$ shows the maximum likelihood estimate under erroneous captures of size γ. The second row is the total population size estimate $\hat{N}_{(\gamma)}$. The third row is the relative bias of $\hat{N}_{(\gamma)}$ to $\hat{N}$.

TABLE 3.12
Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with a Polish nationality,when adjusting for γ erroneous captures.

γ	1	0.9	0.8	0.7	0.6
$\hat{m}_{00(\gamma)}$	157,340	141,596	125,853	110,109	94,366
$\hat{N}_{(\gamma)}$	201,552	185,642	169,733	153,824	137,914
$\hat{N}/\hat{N}_{(\gamma)}$	1	1.09	1.19	1.31	1.46

γ	0.5	0.4	0.3	0.2	0.1
$\hat{m}_{00(\gamma)}$	78,717	62,974	47,230	31,487	15,743
$\hat{N}_{(\gamma)}$	122,100	106,190	90,281	74,372	58,462
$\hat{N}/\hat{N}_{(\gamma)}$ 1.65	1.90	2.23	2.71	3.45	

3.5.3 Tables for section 3.3.3

Table 3.13 shows the sensitivity analysis for the individuals with a Polish nationality by the three registers. The rows display the percentage that has been taken from the CSR. The columns show how much of the percentage taken from the CSR has been linked to either the PR or the ER.

TABLE 3.13
Resulting population size estimates when percentages linkage errors are taken from the CSR (rows) and they are divided differently over the PR and ER (columns) for the Polish individuals.

Percentage in PR→	10	20	30	40	50	60	70	80	90
Percentage in ER →	90	80	70	60	50	40	30	20	10
Percentage from CSR									
10	188,054	195,671	204,215	211,794	222,621	232,873	245,087	258,598	273,813
20	122,628	129,366	137,900	148,190	159,065	171,643	186,595	205,168	226,379
30	84,768	90,377	97,674	105,885	116,075	127,696	142,554	161,835	186,264
40	55,395	64,623	70,021	76,776	84,787	95,314	108,542	125,916	150,558
50	46,185	45,966	54,906	54,723	54,544	70,060	81,068	96,087	118,808
60	29,404	30,769	35,100	39,108	43,869	54,047	53,828	70,987	89,733
70	19,542	22,000	23,439	26,110	29,520	34,019	40,220	53,220	63,912
80	11,684	12,730	14,074	15,668	17,805	20,644	24,625	30,536	40,534
90	5,280	5,754	6,366	7,132	8,146	9,477	11,372	14,290	19,259

Table 3.14 shows the sensitivity analysis for the individuals with an Afghan, Iraqi and Iranian nationality by the

three registers. The rows display the percentage that has been taken from the CSR. The columns show how much of the percentage taken from the CSR has been linked to either the PR or the ER.

TABLE 3.14
Resulting population size estimates when percentages linkage errors are taken from the CSR (rows) and they are divided differently over the PR and ER (columns) for the Afghan, Iraqi and Iranian individuals.

Percentage in PR →	10	20	30	40	50	60	70	80	90
Percentage in ER →	90	80	70	60	50	40	30	20	10
Percentage from CSR									
10	7,695	7,686	7,671	7,657	7,647	7,632	7,623	7,609	7,594
20	6,813	6,788	6,766	6,745	6,724	6,699	6,678	6,658	6,637
30	3,024	3,185	3,340	5,868	5,843	5,814	5,785	5,757	5,729
40	2,190	2,377	2,547	2,733	4,973	4,942	4,909	4,880	4,847
50	1,488	1,675	1,856	2,042	2,235	4,061	4,030	3,998	3,967
60	923	1,104	1,285	1,455	1,635	3,223	3,194	3,164	3,134
70	495	651	810	965	1,117	2,402	2,375	2,350	2,324
80	133	214	295	376	457	1,595	1,575	1,556	1,537
90	39	133	224	314	405	495	585	693	769

3.5.4 Tables for section 3.3.4

Table 3.15 shows the robustness analysis for the Afghan, Iraqi and Iranian people considering erroneous captures of size γ in the CSR register. Row $\hat{m}_{00(\gamma)}$ shows the maximum likelihood estimate under erroneous captures of size γ. The second row is the total population size estimate $\hat{N}_{(\gamma)}$. The third row is the relative bias of $\hat{N}_{(\gamma)}$ to $\hat{N}$.

TABLE 3.15
Erroneous captures for three registers for the Afghan, Iraqi and Iranian individuals

γ	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
$\hat{m}_{00(\gamma)}$	7,249	6,531	5,790	5,072	4,354	3,613	2,895	2,177	1,459	718
$\hat{N}_{(\gamma)}$	68,682	67,932	67,160	66,411	65,662	64,889	64,140	63,391	62,642	61,869
$\hat{N}/\hat{N}_{(\gamma)}$	1.00	1.01	1.02	1.03	1.05	1.06	1.07	1.08	1.10	1.11

TABLE 3.16
Erroneous capture for three registers for the Polish individuals

γ	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
$\hat{m}_{00(\gamma)}$	326,327	293,671	261,016	228,360	195,705	163,049	130,622	97,966	65,311	32,655
$\hat{N}_{(\gamma)}$	450,945	418,146	385,348	352,549	319,751	286,953	254,383	221,584	188,786	155,987
$\hat{N}/\hat{N}_{(\gamma)}$	1.00	1.08	1.17	1.28	1.41	1.57	1.77	2.04	2.39	2.89

4

Different methods to complete datasets used for estimation

Susanna C Gerritse

Statistics Netherlands

Bart F.M. Bakker

Statistics Netherlands/VU University

Peter G.M. van der Heijden

Utrecht University/University of Southampton

CONTENTS

4.1 Introduction

In this manuscript we are interested in the estimation of the population size of the so-called usual residents in the Netherlands. According to the European Union, Regulation (EU) No 1260/2013 of the European Parliament, usual residence is defined as "The place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage". An individual is considered a usual resident when they have lived in the Netherlands for a continuous period of 12 months before the reference time, or if they arrived in the 12 months before the reference time and intend to stay for at least a year. When these circumstances can not be established, "usual residence" means the place of registered residence. The registers used in this manuscript however may register a form of length of stay based on registration date, but not an intent to stay. Hence, only a length of stay may be used to assess usual residence.

The Netherlands has the advantage of having a population register (PR), wherein all registered individuals are documented. Even though for a large part the PR entails the usual residents, for obvious reasons part of the usual residents will be missed by the PR. This incompleteness of the PR has more than one reason. First, within the European Union there is free movement and employment for individuals with a European Union nationality. Usual residents with a European Union nationality do have to register themselves in the PR. Specific rights and services can be provided only to individuals officially registered in the PR. However, unregistered individuals might have forgotten to register, do not know they need to register, or do not want to. Second, the PR is also incomplete due to immigrants, coming from outside the European Union without a working or residence permit. Additionally, as the actual population of the Dutch Census is restricted to those who are registered

This chapter has been published as: S. C. Gerritse, B. F. M. Bakker, and P. G. M. van der Heijden. Different methods to complete datasets used for capture-recapture estimation: estimating the number of usual residents in the Netherlands. Statistical Journal of the IAOS, 31, 613627. 2015. Acknowledgement of author contributions: Theoretical development by BB, SG and PvdH. Document written by SG (main document) and BB (part on previous research), edited by PvdH and BB. Analyses carried out by SG. The authors gratefully acknowledge support in programming by Peter-Paul de Wolf and Maarten Cruyff.

in the PR regardless their residence duration, the estimated true number of usual residents minus the registered population is an indicator of its under count. The registered population will also contain an over count. An over count may occur when registered individuals no longer reside in the Netherlands because of emigration, because they passed away or in the case of administrative delay. In the Netherlands however this is not as big a problem as the under count. Bakker (2009) estimated an over count of 31 thousand individuals, which is only 0.2 percent of the Dutch population.

Because the PR alone is not sufficient to determine the number of usual residents, we linked the PR to an Employment Register (ER) and a Crime Suspects Register (CSR). This enables us to use capture-recapture methodology to assess the part of the population missed by all three registers (Fienberg, 1972; Bishop et al., 1975; Cormack, 1989; International Working Group for Disease Monitoring and Forecasting, 1995). However, because we are interested only in the usual residents of this population, the statistical problem is more complicated. For the PR and the ER residence duration can be derived. However, in the CSR there is no information on residence duration at all. Part of this lack of information in the CSR is solved because this register is linked to the PR and the ER. For the CSR individuals not linked to the PR and/or the ER the information on residence duration is missing.

Therefore we are dealing not only with estimating a population size, but also with handling missing data, since the covariate usual residence is partially missing. Partially missing covariates are usually ignored in capture-recapture problems because they lead to missing data in one or more registers. However, because the covariate usual residence is central in our research question, we cannot ignore it, and we have to solve this missing data problem before we can estimate the population size using capture-recapture methodology.

In estimating usual residence we are interested in what method handles our missing data problem best. When missingness is Missing At Random (MAR) the Expectation Maximization (EM) algorithm can be used to handle the missing data problem (Dempster et al., 1977). In previous research on capture-recapture problems with missing covariates the EM algorithm has been used (Zwane and Van der Heijden, 2007; Sutherland et al., 2007; Van der Heijden et al., 2012; Gerritse et al., 2015b). In these contributions the data are coded into a contingency table format for which the missing data problem is solved by methods developed by Little and Rubin (2002) and Schafer (1997c). See also Fienberg and Manrique-Vallier (2008) for a discussion on the EM algorithm in capture-recapture methods.

However, only part of the information in the observed data seems relevant for solving the missing data problem, as we have great reservations of using the information from individuals that are observed in the PR. The reason is that it is unlikely that individuals that are registered in the PR are relevant for the CSR registered individuals that do not link to the PR or ER. Large part of the individuals in the CSR that are not registered in the PR and do not work in the Netherlands, are assumed to stay in the Netherlands for only a short period; on the other hand, in the PR almost all individuals reside longer than a year in the Netherlands. From the foreign individuals not in the PR but registered in the ER, only 30 percent are usual residents. If only 30 percent of the individuals registered in the ER, but not the PR, are usual residents, compared to nearly 100 percent of the PR registered individuals, it seems implausible to assume that the CSR registered individuals that are not in the PR and ER resemble PR registered individuals. Then the ER registered individuals provide better information about the missing observations in the CSR on the variable usual residence. We note, though, that the EM algorithm is not flexible enough to use only a subpopulation of the data for solving the missing data problem. However, given the missing data is unlikely to resemble the observed data, but only a part of it, the EM algorithm may give biased results.

Another method to handle missing data is multiple imputation. In the context of in capture-recapture analyses we know of one application of multiple imputation used before to impute missing values (Zwane and Van der Heijden, 2007), as well as more general imputation methods (ONS, 2012a). For categorical data Kropko et al. (2014) found that conditional multiple imputation, such as PMM, gave more accurate results in a simulation study on categorical missing data compared to multiple imputation from a joint distribution. In a comparison of imputation methods for binary data (not including EM), PMM outperformed the other imputation methods (Peeters et al., 2015). In this paper we use PMM De Waal et al. (2011); Buuren (2012) as a method to be compared to the EM algorithm. PMM is a sequential multiple imputation method. When data are missing PMM enables the researcher to search the data for a unit that has the same characteristics as the unit that is to be imputed (De Waal et al., 2011). The advantage of PMM is that it provides the researcher with the possibility to use only part of the observed data set as donor to impute the missing usual residence. Thus, where EM has the drawback that it has to use the complete data to impute the missing information, PMM is able to solve the missing data problem in a more appropriate way when only part of the data is relevant for the missing data.

This paper contributes by comparing the EM algorithm and PMM to handle partially missing covariates in capture-recapture analyses. Four scenarios were identified. First, the EM algorithm with a so-called maximal loglinear model will be used to complete usual residence in the CSR. In this scenario, the maximal model is used for both EM algorithm and capture-recapture estimation. Scenario 2 and 3 are used to explore different models for the EM algorithm and the capture-recapture analysis. In scenario 2 the data are completed via the EM algorithm under the maximal loglinear model, comparable to scenario 1. However, capture-recapture is now used on this completed dataset to select the best fitting model. Scenario 3 was used to select the best fitting loglinear model for both the EM algorithm and capture-recapture. Whereas in scenario 2 the loglinear model for EM and capture-recapture will be different, here the restrictive model was kept constant for both completion via EM algorithm and estimation via capture-recapture. The fourth scenario will use the PMM imputation to impute the missing residence duration and will use only ER registered individuals that are not in the PR as donors. Then capture-recapture analysis was carried out on the completed data set.

We continue as follows. In section 2 the data sources used in this manuscript and the linkage process will be explained. In section 3 we will present the results from previous research on the size of usual residents. In section 4 we describe the methods to complete the data and the estimation of the population size in more detail. In section 5 we will present the results, and discuss which scenario gives the best estimate of the usual residents missed by the population register. In section 6 we will conclude this manuscript.

4.2 Data sources and their linkage

Our capture-recapture analysis makes use of three linked registers. The PR is the official Dutch Population Register, in which individuals actively have to register themselves. The ER is a register not documenting individuals but documenting jobs. For the purpose of our analyses the job-register of 2010 has been transformed into a register on individuals. Jobs were attributed to the individuals holding those jobs. Moreover, if a job started in 2010 or was ended in 2010, the jobs are registered with a starting and/or an ending date. The CSR is a register in which suspects of crimes of which the police makes a report are recorded. This register is event based: the units are the reports of the police in which one or more crimes are recorded. There is little information in the CSR and the ER on individuals with ages under 15 and over 65, individuals under 12 can not be registered in the CSR and the ER only registers between 15 and 65. Thus the population specified in this paper consists of the population aged 15 to 65.

The data have been linked for the most part deterministically using a personal identification code that is widely used in Dutch registers. Probabilistic linkage was used to improve this linkage. During linkage it was found that 38 % of the units that were registered only in the CSR had missing information in the linkage variables and therefore were difficult to link to the PR and ER. There is a chance that this group consists mostly of individuals that were either tourists or criminals entering the Netherlands for a short period. This would mean that these individuals are erroneous captures because they do not belong to the population of Dutch residents. It is important to assess whether these individuals substantially affect the population size estimate. However, this research topic is not in the scope of this article and is discussed elsewhere (Bakker et al., 24 - 24 november 2014). In this manuscript we have eliminated a sample of 30 percent of the individuals in the CSR population that did not link to the PR or the ER, assuming that these individuals are erroneous captures. Of these 30 percent 80 percent were individuals that had missing values in the linkage variables and 20 percent were individuals that did not have missing values in the linkage variables. This distribution was chosen assuming that individuals with missing values in the linkage variables had a higher chance of not belonging to the population.

Neither of the three registers has a covariate directly measuring residence duration. However, for two of the three registers we can derive residence duration from information available in those registers. In the PR data are available on the date of registration. In the ER there are data available on joblength. For more details on how the ER residence duration was derived, see Bakker et al. (24 - 24 november 2014). For those individuals in the CSR that link to either the PR, the ER or to both we use the residence duration from the PR or the ER. When residence duration is available from both the PR and the ER the longer residence duration is assumed superior over the residence duration of one of them. In the CSR only there are no variables available to derive residence duration from.

The three linked registers were analysed with loglinear models, as is the standard approach in capture-recapture of human populations, compare (International Working Group for Disease Monitoring and Forecasting, 1995). Four covariates were used in the loglinear models: nationality group, age, sex and usual residence. Initially nationality group has 8 categories : (1) EU15 (excl. Netherlands) (2) Polish (3) Other EU (4) Other western (5) Turkish, Moroccan, Surinam (6) Iraqi, Iranian, Afghan, asylum seeker countries Africa (7) Other Balkan, former Soviet Union, other Asian, Latin American, and (8) other nationalities, not mentioned elsewhere. The countries are clustered according to likely migration motives, migration legislation, regulations of the PR and size. However, in the analysis the last nationality group gave numerically unstable results, and therefore the last two nationality groups were taken together, resulting in 7 nationality groups. For age, we use four levels: (1) 15-24 (2) 25-34 (3) 35-49 and (4) 50-64 years of age.

TABLE 4.1
Observed values for the three registers.

PR	ER	CSR		Total
		Yes	No	
Yes	Yes	2,115	259,804	261,919
	No	4,862	350,551	355,413
No	Yes	355	112,529	112,884
	No	3,561	0	3,561
	Total	12,419	722,884	735,303

Table 5.1 shows the counts for the individuals in the three linked registers ignoring the distribution over the four covariates. The zero count is a structural zero as it represents the number of individuals that belong to the population but are not registered in any of the three registers. One of the aims of the analysis is to find an estimate for this cell and for this purpose a capture-recapture analysis will be executed. Table 1 shows that for the CSR there are more individuals not present than present. In particular the number of individuals in the ER and CSR, but not in the PR is small (355), much smaller than, for example, the number of individuals that are only in the PR (355,413). Given that these 355 individuals are distributed over four covariates, there are many small cell counts in our data, including observed zeros.

4.3 Previous findings

There is previous research on the estimation of population sizes (most notably, Hoogteijling (2002), Bakker (2009) and Van der Heijden et al. (2011)) that overlaps with the population of usual residents that we study in this paper, and these estimates are able to place the estimates found in our scenarios in perspective. However, this previous research shows a wide variety of estimated population sizes depending on the definitions of the population and the methods used, and therefore these studies cannot be used as a simple benchmark for judging the outcomes of our scenarios. Table 4.2 shows their estimates on individuals not registered in the PR.

Hoogteijling (2002) collected different estimates from earlier research in the nineties. In order to achieve an estimate of the size of the population not registered in the PR and living four months or longer in the Netherlands in 2000, she combined the available information from different sources. Neglecting some very small categories, the population can be estimated by adding illegal immigrants, adding the balance of wrongfully not registered residents and wrongfully registered non-residents, and recently arrived asylum seekers who have not registered because they are not allowed to do so yet. This results in an estimate of 73 to 149 thousand missed residents, with a mean of 111 thousand, being less than 1% of the registered population (Table 4.2).

Bakker (2009) also used information from different sources to get an estimate of the under and over coverage of the PR in 2006, having the same definition of usual residence as Hoogteijling (2002), so those who stay longer than four month in the Netherlands are supposed to be usual residents. He distinguishes the different categories of which it is known that they are missed or are over counted in the PR and he estimates their numbers with different sources. He estimates the total under count as 205 thousand usual residents. However, there is a large uncertainty because some of the estimates are quite arbitrary. The largest contribution is from illegal immigrants whose size is estimated between 74 and 184 thousand. The total number of missed persons is 236 thousand, where 31 thousand persons are still in the population register while they have left the country or have died.

Van der Heijden et al. (2011) used capture-recapture methodology to estimate the missed portion of the population in 2009 from Poland, Bulgaria, Romania and other nationality groups in middle and eastern Europe new in the EU (i.e. Hungary, Czech republic, Slovakia, Slovenia, Latvia, Lithuania and Estonia). Therefore, their outcomes can only be compared to two nationality groups used in this paper. They used the PR and the CSR as sources and applied capture-recapture methods to estimate usual residents in the same definition as we do. A difference between Van der Heijden et al. (2011) and this manuscript is that in this manuscript assumed erroneous captures have been excluded from the analysis, whereas in Van der Heijden et al. (2011) these may still have impact on the estimate. Additionally, the number of usual residents is given for the total population of individuals with a middle and eastern

European nationality from new EU countries residing in the Netherlands, including those registered in the PR. There are 200 thousand usual residents with a middle and eastern European nationality not registered in the PR.

TABLE 4.2
Overview of previous research to individuals residing in the Netherlands.

		min.	max.	Total/mean	of which usual residents	
		x1000			%	x1000
Hoogteijling (2002), estimates for the year 2000						
	Registered population			15987		
plus	illegal immigrants	46	116	81	80	65
balance	wrongfully not registered residents and wrongfully registered non-residents	-15	-16	-15.5	80	-12
plus	asylum seekers not yet registered as residents	42	49	45.5	80	36
	missed population of residents (≥4 month)	73	149	111		89
	Total population of residents (≥4 month)	16,060	16,136	16,098		16,076
Bakker (2009), estimates for the year 2006						
	Registered population		16,334			
plus	illegal immigrants	74	184	129	80	103
plus	foreign labour force			64	30	19
plus	foreign students			25	30	8
plus	diplomats and NATO military			6	80	5
plus	asylum seekers not yet registered as residents			6	80	5
balance	administrative delay			5	80	4
minus	non-residents working abroad temporarily			-29	30	-9
	missed population of residents (≥4 month)			205		135
	Total population of residents (≥4 month)			16,509		16,469

It is difficult to describe the expected value of the size of the population of usual residents in 2010, because some estimations are outdated and some do not use the same definition, or both. However, by harmonizing the results for the definitional differences and looking at the developments of the number of new asylum seekers and the number of foreign workers, we can provide a range of expected outcomes. These expected outcomes could help in providing a perspective where the current estimate of usual residents may be compared to.

In the under count of 111 thousand found by Hoogteijling (2002) the majority is former or present asylum seeker. Because the procedures for seeking asylum had a long duration, certainly with a mean longer than a year, we assume that most residents who were not registered as such stayed for longer than a year in the Netherlands. Therefore we assume 80 percent of the 111 thousand not registered to be usual residents, which comes down to 89 thousand usual residents in 2000.

Bakker (2009) estimated an over count of 205 thousand and this estimate is difficult to harmonize with the definition of usual residence in this manuscript, because we do not have empirical information on the residence duration of the different categories that are over counted in the PR. However, if we assume (i) that 80% of the illegal immigrants are a usual resident because they still are in majority former asylum seekers and (ii) that the same percentage is true for smaller categories like asylum seekers, diplomats and NATO military and administrative delay of new born and immigration, and (iii) that 30% of the foreign work force and foreign students is a usual resident, the same percentage as we found for the foreign work force in 2010 (Bakker et al., 24 - 24 november 2014), then the estimated number of usual residents not registered in the PR is 135 thousand in 2006.

The estimate of Van der Heijden et al. (2011) for the usual residents in 2009 from Eastern European countries uses the same definition and does not have to be harmonized. However, as they did not adjust their estimation for erroneous captures, the estimate of 200 thousand is expected to be too high. Bakker et al. (24 - 24 november 2014) show a decrease of approximately 37% if they correct for erroneous captures, we expect that the estimation of the size of the usual residents would be 126 thousand only from Eastern Europe.

Two significant developments have to be mentioned to explain changes in the number of not registered usual residents between 2000 and 2010. The first is the decline of the number of asylum seekers between 2000 and 2010 (Figure 1). The numbers dropped from almost 45 thousand in 2000 to 10 thousand in 2004, among else due to changed regulations. After 2004 there is a more or less constant number of asylum requests between 7 and 15 thousand. The other one is the sharp rise

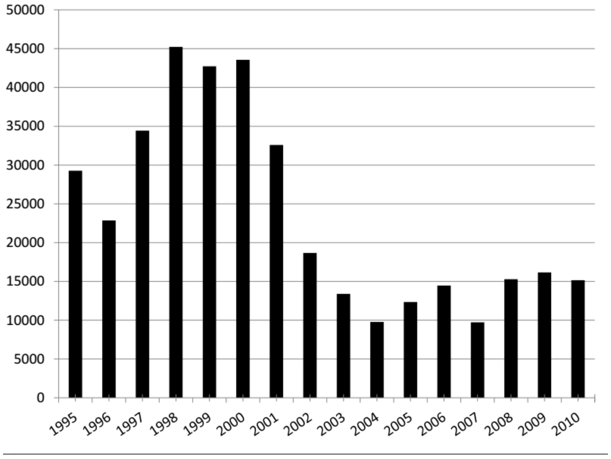

FIGURE 4.1
Number of asylum requests in the Netherlands 1995-2010 (Central Bureau of Statistics)

of the foreign workforce from the year 2006, in particular from Eastern Europe, who did not register themselves in the population register. This was in 2006 121 thousand and increased to 182 thousand in 2010 (CBS, StatLine, 2015). This development was possible because the civilians of these countries could enter the Netherlands without a residence permit and after 2007 for the most part could also work without a working permit.

We arrive at the following conclusion, cautiously indicating that it is always dangerous to extrapolate earlier estimates to later periods. We expect that the number of usual residents not registered in the PR has been increased since the year 2000 to 175 to 225 thousand. The total number is certainly much higher than the 135 thousand in 2006 because of the inflow of migrant workers from Eastern Europe since then. On the other hand, the number of asylum seekers has been constant since 2006 and will not cause important developments. If the estimation of the number of not registered usual residents from Eastern Europe in 2009 is correct, then it is reasonable to assume that the upper bound is approximately 225 thousand, because the 100 thousand not registered usual residents from other countries will not have disappeared.

As can be seen from Table 5.1 there are 116,445 registered individuals not in the PR but in the CSR and/or the ER. Of these 116 thousand individuals, we found that 33 thousand are usual residents, and thus are part of the known under coverage, these individuals have to be added to the estimate from the scenarios (Bakker et al., 24 - 24 november 2014).

4.4 Methods

We will use different scenarios for handling the missing data and estimating the part of the population missed by all three registers. Here, we describe the scenarios and evaluate them in the context of our problem. We will first give a short introduction to capture-recapture analysis using loglinear modelling and then we will discuss the EM algorithm and multiple imputation in this context.

4.4.1 Capture-recapture methodology using loglinear modelling

For estimating the size of human populations loglinear modelling seems to be the most popular method. It was discussed in depth in the standard work by Bishop et al. (1975) and since then it has been reviewed regularly, for example by Cormack (1989), the International Working Group for Disease Monitoring and Forecasting (1995) and Chao et al. (2001).

The simplest loglinear model for estimating the size of a population is based on two linked registers, A and B. Let the levels of A be indexed by i (i = 0,1) where $i = 0$ stands for ”not included in register A”, and $i = 1$, stands for ”included in register A”. Similarly, let the levels of B be indexed by j ($j = 0, 1$). Expected values are denoted by m_{ij}. Observed values are denoted by n_{ij} with $n_{00} = 0$, because there are no observations for the cases that belong to the population but were not present in either of the registers.

After linkage there is an observed number of individuals both in A as well as in B, n_{11}, an observed number of individuals only in A but not in B, n_{10} and an observed number of individuals only in B and not in A, n_{01}. Individuals being neither in A nor B are missed and capture-recapture can estimate this missing number, where we denote this estimate by $\hat{m}_{00}$. Assuming statistical independence of being in A and being in B, the odds ratio between being in A and being in B is 1, i.e. $n_{11}\hat{m}_{00}/n_{10}n_{01} = 1$. It follows that $\hat{m}_{00} = n_{10}n_{01}/n_{11}$. Then the population size N is estimated as $\hat{N} = n + \hat{m}_{00}$, where n is the observed number of individuals, i.e. $n = n_{11} + n_{10} + n_{01}$. The link between these equations and loglinear modelling is that loglinear parameters are functions of odds ratios. In the loglinear model for two variables, assuming that the odds ratio is 1 comes to the same as assuming that the interaction parameter between A and B is absent. The independence model just described is denoted in loglinear model notation as [A][B], showing that being in register A is unrelated to being in register B.

By including a third register C the assumption of statistical independence is replaced by the assumption that there is no three factor interaction. This model is denoted by [AB][AC][BC]. In other words, there may be interaction between A and B, but this interaction is identical in the sub-tables for individuals included in C and not in C. This also suggests a way to find an estimate $\hat{m}_{000}$: as the odds ratios are identical, $n_{111}n_{001}/n_{101}n_{011} = n_{110}\hat{m}_{000}/n_{100}n_{010}$, and this can be solved for $\hat{m}_{000}$.

Categorical covariates that are available in each of the registers, such as age and sex, can be easily added. If we collect them in a stacked variable X, the model becomes $[XAB][XAC][XBC]$. This model is saturated as the number of parameters is identical to the number of observed counts, and fitted values are equal to observed counts.

It is a crucial part of the capture-recapture procedure to search for a model that is more parsimonious than $[XAB][XAC][XBC]$, yet fits the data well. Parsimonious models have the advantage that the resulting estimate of N is more stable (i.e. has a smaller confidence interval) than saturated models (Agresti, 2013). The fit is usually evaluated using the deviance, that follows a chi-squared distribution with the number of counts minus the number of independent parameters when the model is true. Here the number of counts is (2*2*2 -1)*(number of levels of the stacked covariates), where the 1 refers to the cell that has a count of zero by design, hence this cell is called a structural zero. The number of possible models is large and there often is lack of theory that points out which models are of particular interest. Therefore exploratory model searches are usually employed, such as forward selection, backward elimination and stepwise procedures, that we also know from linear multiple regression. Using the difference deviance test, the AIC or the BIC a final model is chosen, where a model with either the lowest AIC or the lowest BIC is preferred. In contrast to the difference deviance test the AIC and the BIC can also be used to choose between non-nested models. Both the AIC as well as the BIC are a function of the deviance. The BIC leads to more parsimonious models, because the AIC has a penalty of $2k$ but the BIC a penalty of k log n, where k is the number of free parameters and n the observed count. We use the BIC, because in capture-recapture problems with a large observed count n, we prefer the BIC to prevent over fitting.

4.4.2 Missing covariate

The capture-recapture problem becomes more complicated when there are covariates involved that are not observed in every register. In the data that we study in this paper this holds for usual residence, which is not observed in the police register CSR. We consider this to be a missing data problem: the variable usual residence is missing for those individuals who are only in the register CSR. This type of missing data problem has been worked out in detail for two registers, see (Zwane and Van der Heijden, 2007; Van der Heijden et al., 2012; Gerritse et al., 2015b). Here we have a capture-recapture problem with three registers instead of two, so this case deserves careful attention. Table 4.3 illustrates.

TABLE 4.3
The Polish individuals by the three registers and usual residence. The two missing cells add up to 1,043.

UR	PR	ER	CSR	
			Yes	No
No	Yes	Yes	32	3,523
		No	34	3,225
	No	Yes	149	60,190
		No	missing	0
Yes	Yes	Yes	183	21,309
		No	195	14,052
	No	Yes	81	20,216
		No	missing	0

Table 4.3 shows the cross-classified counts for the three registers PR, ER and CSR, split out by UR and non-UR, for Polish individuals. Notice that there are 16 cells in total, where two are structurally zero (they refer to the individuals that are missed by all three registers), and for two cells we only know the sum, namely for not in PR, not in ER, but in CSR. Only the sum is known because for these individuals usual residence is missing, so we cannot split them up over the cells yet. This holds for 1,043 individuals, as indicated in the header of the table.

To simplify the discussion of loglinear models, we now use the variables P for PR, E for ER and C for CSR, and U for usual residence. If usual residence would not be missing, the saturated model with as many parameters as counts would be $[PEU][PCU][ECU]$. The saturated model has 14 parameters, and it is saturated with the property that the fitted counts are identical to the observed counts. However, due to the missingness of the variable usual residence in the CSR there is one count less (the two missing counts add up to 1,043) and the maximal model for these data only has 13 parameters. The distinction between a so-called maximal model and the standard saturated model is that in the maximal model certain parameters are zero by design. The parameter for the interaction between P, C and U can be estimated from the counts 32, 3,523, 149, 60,190, 183, 21,309, 81 and 20216; the parameter for the interaction between P, C and U can be estimated from 32, 3,523, 34, 3,225, 183, 21,309, 195 and 14,052; but the parameter for the interaction between P, E and U is not identified because for the 1,043 individuals their level on U is not identified. Therefore the maximal model becomes $[PE][PCU][ECU]$. Interestingly, because of the identification problem the model has no parameter for the interaction between P, E and U. When individuals are not in P and E but only in C, the data for usual residence are missing by design. As a result the information on U in the maximal loglinear model is only available in the margins of the variables P, C and U, and the margin of E, C and U, but not in the margin of P, E and U. For more results on maximal models, see (Zwane and Van der Heijden, 2007).

For evaluating scenarios it is important to note that model $[PE][PCU][ECU]$ makes two assumptions. First the three-factor interaction between P, E and C is zero. In other words, the relation between P and E is identical for those who are in C and those who are not in C (and, similar statements can be made for the relation between P and C and between E and C). Secondly, the interaction between P, E and U is zero, meaning that the interaction between P and E is identical for those individuals who have a usual residence shorter than a year and those who have a usual residence longer than a year. However, this last assumption is not very plausible. It is known that those who stay longer in the country, in particular from Eastern Europe, assimilate fast in society, find permanent work and a partner (Gijsberts and Lubbers, 2015). Therefore they will register themselves more frequently than those who only live in the Netherlands for a short period. In other words, it is plausible that for those who reside in the Netherlands for longer than a year the odds ratio between P and E will be larger than for those who reside in the Netherlands for shorter than a year. Yet the maximal model cannot accommodate this because the interaction between P, E and U cannot be estimated.

These preliminaries bring us to the definition of the scenarios. Our statistical problem has two aspects:

(i) there are missing data on the variable usual residence, and

(ii) using capture-recapture methodology we are going to fit a loglinear model under which the part of the population that is missed by all three registers is estimated.

Our scenarios differ in the way that the two steps are taken. We make use of two procedures for handling the missing data problem: the Expectation-Maximization method (EM) and multiple imputation using Predictive Mean Matching (PMM).

4.4.3 Scenarios using the EM algorithm

The EM algorithm is a general iterative algorithm for maximum likelihood estimation when data are incomplete (Little and Rubin, 2002). The EM algorithm consists of an Expectation (E) step and Maximization (M) step. In general in the E step the algorithm replaces missing values by values that are expected under a given model. Then under the M step the algorithm estimates parameters that are maximized on the expected values of the E step. Then in the next E step expectations are calculated for the missing values using the current best parameter estimates found in the last M step, after which a new M step maximizes the parameters using the data completed in the E-step. This is repeated until convergence occurs, where the joint distribution of the register and covariate variables are preserved, under the loglinear model specified. Hence MAR is assumed under the joint distribution given loglinear modeling. For the maximal model, convergence is after only one iteration.

After completion of the EM algorithm, the completed data can be used for capture recapture estimation. When the loglinear model for capture-recapture estimation is identical to the loglinear model the parameter estimates from the EM algorithm can be used to estimate the population size. Then the EM algorithm alone suffices to estimate the missed portion of the population. However, when after EM completion another loglinear model is preferred for capture-recapture analysis, the EM completed data can be used as input for the capture recapture analysis.

We distinguish the following scenarios. For scenario 1 we use the EM algorithm where the loglinear model chosen is the maximal model. For capture recapture analysis the maximal model is also used, and thus the parameter estimates from the EM algorithm can be used to estimate the missed portion of the population. In scenario 2 we also use the maximal model for the EM algorithm to complete the data. However, this completed data are then used as input for capture-recapture analysis where the function STEP in R is used to look for the best fitting, parsimonious loglinear model to the completed data. Then the EM algorithm and capture-recapture are done in 2 steps. In scenario 3, just as in scenario 1, we use the parameters from the EM algorithm to estimate the missed portion of the population. However, unlike scenario 1, scenario 3 will use more restrictive models for the EM algorithm.

4.4.4 Scenario for multiple imputation using predictive mean matching

A fourth scenario is multiple imputation using Predictive Mean Matching (PMM). When an individual has missing data PMM enables the researcher to search in the data for individuals that have the same characteristics as the individual that has values that need to be imputed, and use their observed values to impute the missing value (De Waal et al., 2011). MAR is assumed, in that units with the same background characteristics will have similar values on the missing variable, if this variable would have been observed.

Predictive mean matching is an example of a hot deck, nearest neighbour multiple imputation method. Missing values are imputed using values from the complete cases matched with respect to some metric. All individuals with the same background variables as the missing value are candidate donors for imputing. From these donors, one random donor is sampled from the candidates and the value on usual residence from this donor candidate is taken as a value for the missing unit (Buuren, 2012; Little and Rubin, 2002). By selecting individuals from the same background values, the joint distribution based on the background variables is preserved.

The PMM procedure has been repeated ten times, to account for the uncertainty of the individual imputations. To estimate the number of usual residents, the capture-recapture method has been applied to all imputed datasets. To estimate the number of usual residents, the mean of the ten estimates has been computed.

PMM has the advantage that it allows to select a specific subpopulation for which it can be assumed that it resembles the subpopulation that has to be imputed best. In this case, we have to find a donor population for the CSR-records that do not link to the PR and ER. In this donor population it is presumable that the residence duration is relatively short, because there is a positive association between residence duration and registration. Therefore, for the donors we choose individuals in the ER that do not link to the PR, i.e. individuals who are working but did not register as a resident of the Netherlands, thus we assume MAR in a conditional distribution. They also have a relatively short residence duration because of this aforementioned association.

4.4.5 Concluding remarks

We summarize the scenarios here. Both the EM algorithm and multiple imputation using PMM are established methods with a solidly grounded base in literature. Both methods assume Missing At Random (MAR). There are also two important differences.

First, the EM algorithm cannot make use of models that are more complicated than the maximal model. In the maximal model the interaction between P, E and U cannot be estimated. For the EM algorithm the missing data are completed from a joint distribution of the observed data under a given loglinear model. It has been argued above that this is a drawback for our missing data, which is assumed to resemble only a subpopulation of the observed data. On the other hand, PMM is applied using as a donor population the subpopulation of individuals that are in ER but not in PR. For this subpopulation the relation with usual residence is used.

Second, both methods handle missing data differently. EM algorithm completes the incomplete data according the loglinear model specified. Note that differences in loglinear models may result in different estimates (Little and Rubin, 2002; Van der Heijden et al., 2012). Thus the choice of the loglinear model is important. Predictive Mean Matching (PMM) is a sequential multiple imputation method. When data are missing PMM enables the researcher to search the data for a unit that has the same characteristics as the unit that is to be imputed (De Waal et al., 2011). The advantage of PMM is that a missing unit will be given the same value on the missing variable of an observed unit. It is assumed that units with the same background characteristics will have similar values on the missing variable, if this variable would have been observed. Then PMM has the advantage of assuming MAR between the missing data, and the observed data that resembles the missing data best. Multiple imputation using PMM is flexible in the sense that it is possible to use only that part of the table that seems most appropriate to use for the problem at hand. So in an evaluation of both differences multiple imputation by PMM seems better suited to handle the problem that we study.

Throughout the paper, the software R has been used for all computations. For the EM algorithm the package CAT (Meng and Rubin, 1991; Schafer, 1997a,b) was used. For multiple imputations using PMM the package MICE was used (Buuren, 2012). After completion, the R-function GLM was used to estimate the missing part of the population. We used parametric bootstrap confidence intervals to estimate a 95% confidence interval for the point estimate of usual residents. 10,000 bootstrap samples are used.

4.5 Results

4.5.1 Scenario 1: Maximal model for EM estimation and capture-recapture analysis

In our first scenario the missing data are completed under the maximal loglinear model, and the missing part of the population is also estimated under this model. This approach is carried out for the nationality groups separately.

For almost all of the nationality groups the population size estimates tend to infinity. To examine why we get these results, we have to take in mind that we employ a two-step process. First the incomplete data is completed via EM algorithm and then the capture-recapture analysis is carried out to get an estimate of the missed portion of the population, and problems can occur in both completion of the data or estimation under capture-recapture analysis.

As an illustration of what goes wrong we go back to Table 4.3, which is a marginal table of individuals with a Polish nationality where we added up over the covariates sex and age. There are two structurally zero cells, representing the part of the population that is not observed, and two cells that are zero for which usual residence has to be estimated, but where the sum should be 1,043.

First the missing values for usual residence is estimated under the maximal model. This yields 655 individuals categorized as residing shorter than a year, and 376 as longer than a year. The resulting table is the input data for the capture-recapture analysis. Estimates are derived again under the maximal model. As in the maximal model fitted counts are equal to observed counts, it is important to realize that the observed counts in Table 3 are further split up over age and sex. Hook and Regal (2000) discussed that some models, especially the saturated model (in our case, the maximal model), is sensitive to small or zero cells. As an example, Table 4.4 shows the 376 out of 1,043 individuals in the CSR only, classified as usual residents. There are only 2 women with an age between 50 and 64, whereas there are 137 young men. Because our data consists of both large and small cell numbers, and in this scenario we used the maximal model for capture-recapture analysis, it follows that the resulting estimate is implausibly high. Capture-recapture analysis thus is sensitive under the maximal model in contingency tables with small and zero cells, which we have. Thus the maximal model cannot be reliably used for capture-recapture analysis under the current data.

TABLE 4.4
Estimated Polish usual residents that are missed by all three registers, by Age and Sex

Age	Men	Women
15 - 24	137	26
25 - 34	138	10
35 - 49	53	3
50 - 64	7	2

4.5.2 Scenario 2: Maximal model for EM estimation, restrictive models for capture-recapture analysis

In this scenario, after the EM algorithm was used for completing the missing data under the maximal model, we used the function STEP in R to choose the best fitting loglinear model via the BIC for the capture-recapture analysis. Table 4.5 shows the results. We find a total of 659 thousand individuals missed by all three registers, 139 thousand of those individuals are usual residents (confidence interval 120 – 176 thousand). The usual residents are 21 percent of the total missed portion of the population. Scenario 2 resulted in parsimonious and more stable models where the outcome seemed more plausible.

The loglinear models for this scenario can be found in the Appendix. The models for scenario 2 are more parsimonious and restrictive than the maximal model from scenario 1. For example, take the model of the individuals with a Polish nationality. The last term in this model [PUSA] is comparable to the first term in the model for scenario 1, which was [PCUSA], but in scenario 2 the data are collapsed over the CSR. In deleting CSR from the interaction term the distribution of individuals over the contingency table becomes more balanced, and this leads to estimates from the capture-recapture analysis that are numerically more stable.

TABLE 4.5
Estimates for scenario 2

Nationality	Total missed	< 1 year	≥ 1 year	95% CI
	x1000	x1000	x1000	x1000
EU15	146	104	42	30 - 47
Polish	265	211	53	49 - 69
Other EU	155	132	23	13 - 30
Other West	16	12	4	3 - 5
Turkey, etc.	3	1	1	.8 - 2
Iraq, etc.	9	8	2	1 - 2
Balkan, etc. Other.	65	51	14	10 - 28
Total	659	520	139	120 - 176

Estimates of the missed portion of the population per nationality for scenario 2, where the maximal model is used for EM algorithm. The analysis was done with the best fitting restrictive loglinear model.

4.5.3 Scenario 3: Restrictive models for both EM estimation and capture-recapture analysis

In scenario 3 we use more restrictive models for the EM algorithm, and keep models for EM and capture-recapture analysis equal. This is the standard approach of using the EM algorithm. Results can be found in Table 4.6. There are 129 thousand usual residents missed (CI is 111–170 thousand), which again is 20 percent on 608 thousand total individuals missed by the three registers. This total estimate is similar to the estimate for scenario 2, however, the estimates per country are somewhat different. Interestingly, as can be seen in the Appendix, the best fitting models for capture-recapture analysis on the completed datasets are quite different between these two scenarios.

Table 4.7 shows the data after completion via EM algorithm under a more restrictive model. Note that the data differ compared to Table 4.3. Under more restrictive models the content of the table can change to maximize the fit of the margins under the loglinear model. Thus the completed table differs from the observed Table 4.3.

4.5.4 Scenario 4: multiple imputation using PMM

Table 4.8 shows that when we use multiple imputation using PMM and conduct a capture-recapture analysis on the data, 610 thousand individuals are missed by all three registers, of which 179 thousand are usual residents(with a larger CI of 121–237 thousand). Now there are 29 percent usual residents on the total number of individuals missed by all three registers, a slight increase compared to scenario 2 and 3. The CI

TABLE 4.6
Estimates for scenario 3

Nationality	Total missed	< 1 year	≥ 1 year	95% CI
	x1000	x1000	x1000	x1000
EU15	147	104	42	31 - 59
Polish	243	195	48	38 - 62
Other EU	142	123	20	13 - 35
Other West	18	13	4	3 - 7
Turkey	3	2	2	1 - 2
Iraq	10	8	2	2 - 3
Balkan, etc, Other.	45	34	11	9 - 13
Total	608	480	129	111 - 170

Estimates of the missed portion of the population per nationality for scenario 3, where the same restrictive loglinear model is used for EM algorithm and capture-recapture analysis.

TABLE 4.7
The data for the Polish individuals after completion with EM algorithm via restrictive loglinear models

UR	PR	ER	CSR	
			Yes	No
No	Yes	Yes	23	3,530
		No	33	3,226
	No	Yes	158	60,180
		No	781	0
Yes	Yes	Yes	193	21,300
		No	196	14,050
	No	Yes	71	20,225
		No	261	0

for usual residents is higher than the estimates resulting from the EM algorithm when more restrictive models are used, such as in scenario 2 and 3.

After PMM imputation the data are similar to Table 4.3 in that the observed part of the data remains unchanged. However, the 1,043 individuals in the CSR are distributed differently per imputation, where generally less usual residents are imputed than for scenarios 2 and 3. As can be seen from the Appendix the loglinear model best fitting this completed data is different from the other three scenarios and a different estimate results.

TABLE 4.8
Estimates for scenario 4

Nationality	Total missed	$<$ 1 year	$\geq$ 1 year	95% CI
	x1000	x1000	x1000	x1000
EU15	156	112	46	24 - 68
Polish	240	178	61	28 - 95
Other EU	138	95	42	16 - 68
Other West	15	11	5	1 - 9
Turkey, etc,	4	2	2	.5 - 4
Iraq, etc,	9	3	6	5 - 7
Balkan, etc, Other.	48	31	17	12 - 22
Total	610	431	179	121 - 237

Estimates of the missed portion of the population per nationality for scenario 4, after PMM imputation and restrictive loglinear modelling capture-recapture analysis.

4.6 Discussion

Four scenarios were defined to assess the effect of different methods of completing missing data on the population size estimate from capture-recapture. Usual residence has been completed in three different ways. For scenario 1 and 2 the maximal loglinear model was used via EM algorithm. The first scenario turned out to be plagued by numerical difficulties and will be further ignored in the discussion. The second scenario yielded a point estimate of 139 usual residents with a 95 percent confidence interval of 120 − 176 thousand. The third scenario also employed the EM algorithm but used more restrictive loglinear models. This scenario led to a point estimate of 129 thousand with a confidence interval of 111–170 thousand. Scenario 4 used multiple imputation by means of PMM. Here the point estimate was 179 thousand with a larger confidence interval of 121 –237 thousand. Therefore the focus of our discussion is on scenario 2 and 3, with a lower point estimates and lower confidence intervals, versus scenario 4, with a higher point estimate and a wider confidence interval.

In Section 3, in the discussion of earlier estimates it became clear that approximately 33 thousand individuals in the ER and CSR that did not link to the PR are usual residents. This number has to be added to the estimate of usual residents missed by all three registers for the scenarios 2, 3 and 4. Then we get 172 thousand for scenario 2, 152 thousand for scenario 3 and 212 thousand for scenario 4. This means that only the estimate of scenario 4, for the PMM imputed data, lies within the range of 157 to 225 thousand that was based on earlier estimates.

Both the EM algorithm and the PMM method are respected methods that work well when the assumptions are fulfilled. When used in an identical context, both methods perform equally well in terms of point estimates and confidence intervals (Buuren, 2012). However, in this application the important difference is that the EM uses the complete data set for deriving information on how to impute usual residence whereas PMM only uses the observations that are not in the PR. We have argued that the approach of PMM has a better theoretical motivation because it is likely that the persons with a missing value for usual residence will be similar to the persons not in the PR. Thus the use of PMM is theoretically better motivated. We note also that the observed number not in the PR is with 116 thousand much smaller than the observed number in the PR, which is 617 thousand. So in PMM the size of the population used for imputation is smaller than the size of the population used in the PR. This leads to the following interpretation: we expect that *on average* PMM will lead to a population size estimate that is closer to the true population size estimate than the EM estimate. However, because the number of observations used in PMM is smaller, the estimated confidence interval is larger.

It is not easy to make a choice between bias and variance. Choosing for the EM algorithm would mean that the point estimate of PMM would be outside the EM confidence interval. So by choosing the EM algorithm it would be likely that the point estimate would be too low. This is a first reason to prefer the PMM estimate. A second reason is that it falls in the range of what we expected based on earlier results. We understand that both of these reasons are not completely convincing and satisfactory. This holds in particular for the second reason, as the range based on earlier research is based on studies that overlap with the research described in this manuscript, but the earlier research differs from the research discussed here in terms of data sources and models used.

We shortly discuss a capture-recapture estimation without using the CSR. When a capture-recapture analysis is conducted on PR and ER alone only 27 thousand individuals are estimated as the missed portion of the population. This estimate is implausibly low, given that half of that number alone is an asylum seeker (See Figure 4.1). In deleting the CSR from all interaction terms we assume being in PR and being in ER are statistically independent, which is not realistic. Moreover, capture-recapture analysis for only PR and ER will result in a very specific population, that does not include illegal immigrants. Hence CSR is important for our capture-recapture analysis, but may be deleted from some interaction terms.

4.7 Conclusion

Often covariates are used in capture-recapture estimation for estimating hard-to-reach populations. When covariates are important for answering a research question, missing data in these covariates can pose a problem. Since such a covariate cannot be left out of the analysis, a solution has to be sought to handle the missing data problem. In this paper the covariate that poses a problem is usual residence. Since we are interested in estimating usual residents we cannot exclude the covariate.

There are multiple ways of handling the missing data for categorical data sets. In this paper we have chosen to compare the EM algorithm and multiple imputation using PMM. For EM algorithm three variants were chosen, (i) a maximal model for EM completion and capture recapture estimation, (ii), a maximal model for EM algorithm and a more restrictive model for capture-recapture analysis, and (iii) a restrictive loglinear model that is identical for EM completion and capture-recapture analysis. After multiple imputation using PMM the best fitting loglinear model is chosen for capture-recapture analysis

Scenario 1 gave unrealistic estimates, scenario 2 and 3 gave lower estimates which are similar to one another, and scenario 4 gave a higher estimate than scenario 2 and 3. The confidence intervals of scenarios 2 and 3 were smaller than that of scenario 4. Theoretically PMM is better motivated to handle missing data in our context because PMM is more flexible in dealing with missing data when a specific subpopulation is missing. Additionally, the estimate after PMM imputation of 212 thousand usual residents not registered in the PR was the only one that lies in the prespecified range. For this reason we have a preference for the estimate of scenario 4, noting that both reasons are not completely convincing and satisfactory.

At the beginning of this manuscript we stated that the registers do not register intent to stay, but only a length of stay. However, we assume the individuals in the PR do have an intent to stay for a longer period. Individuals register themselves in the PR, which by law has to be done when an individual is in the Netherlands for four months or longer, or intents to stay for four months or longer. Given the assumption that a person registers in the PR with the intent to stay for a longer period, we assume all PR registered individuals to intent to stay for longer than a year. Thus all PR registered individuals will be considered usual residents. Given that there are 16,638 thousand individuals registered in the PR, and assuming these individuals have registered to reside for a longer period in the Netherlands and thus are usual residents, the total number of usual residents in the Netherlands is 16,856 thousand, of which 1.3% are not registered in the PR.

4.8 Appendix

Maximal model for scenario 1.
$[PCUSA][ECUSA][EPSA]$

TABLE 4.9
Loglinear models per nationality for scenario 2

Nationality	Model
EU15	[CS] [CE] [SE] [PE] [PCU] [PUS] [PCA] [PUA] [CUA] [PSA] [USA] [USE] [UAE] [SAE]
Polish	[PC] [CE] [PE] [PUS] [CUS] [PUA] [CUA] [PSA] [USA] [USE] [SAE] [PUSA]
OthE EU	[CE] [PE] [PCU] [PCS] [PUS] [CUS] [PCA] [PUA] [CUA] [PSA] [USA] [USE] [UAE] [SAE] [PCUS] [PUSA] [USAE]
OthE West	[PU] [PS] [CS] [PA] [PE] [USA] [USE] [UAE] [SAE]
TUkey etc.	[PE] [PCU] [PCS] [PUA] [CUA] [CSE] [CAE] [USAE]
Iraq, etc.	[PU] [PS] [CS] [US] [PA] [CA] [UA] [CE] [UE] [PE] [SAE]
Balkan, etc. Other.	[[CE][PE][PCU][PCS][PUS][CUS][PCA][PUA][USA][USE] [UAE][SAE][PCUS]

TABLE 4.10
Loglinear models per nationality for scenario 3

Nationality	Model
EU15	[PC] [PE] [PU] [PS] [PA] [CE] [CS] [EU] [ES] [EA] [US] [UA] [SA]
Polish	[PC] [PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA] [SA]
Other EU	[PC] [PE] [PU] [PS] [PA] [CE] [CS] [CA] [EU] [US] [UA] [SA]
Other West	[PU] [PS] [PA] [CE] [CU] [CS] [EU] [ES] [EA] [US] [UA] [SA]
Turkey, etc.	[PC] [PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA] [SA]
Iraq, etc,	[PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA] [SA]
Balkan, etc. Other.	[PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA]

TABLE 4.11
Loglinear models per nationality for scenario 4

Nationality	Models
EU15	[PE] [CS] [CA] [PC] [EAU] [SAU] [ESA] [PAU] [ESU] [PSU] [PSA] [ECU]
Polish	[EC] [CS] [EA] [EU] [CU] [PC] [PE] [CA] [PSA] [PSU] [SAU] [PAU] [PSAU]
Other EU	[EC] [PC] [PSA] [PSU] [PEU] [PEA] [SAU] [PCS]
Other West	[PS] [CS] [PA] [ESA] [EAU] [PEU]
Turkey, etc.	[PC] [ESA] [SAU] [PEU] [ECA] [EAU] [ECS] [PES] [CAU]
Iraq, etc.	[CS] [PE] [CA] [SU] [PS] [PA] [EC] [PEU] [ESA] [EAU]
Balkan,etc. Other	[CS][PS][EC][PA][CA][CU][PEU][EAU][ESA][ESU][SAU][PAU][CAU]

These models are exemplary for one of the ten multiple imputations

5

Undercoverage of the population register in the Netherlands, 2010

Susanna C Gerritse
Statistics Netherlands

Bart F.M. Bakker
Statistics Netherlands/VU University

Peter-Paul de Wolf
Statistics Netherlands

Peter G.M. van der Heijden
Utrecht University/University of Southampton

CONTENTS

5.1 Introduction

In this manuscript we are interested in estimating the under coverage of the Dutch Population Register (PR). The PR is an important register for Census purposes, and the undercoverage of this register gives an indication of the quality of the register. One common method is to link the PR to other registers and estimate the portion of the population missed by the registers using capture-recapture estimation, also known as multiple systems estimation (Fienberg, 1972; Bishop et al., 1975; Cormack, 1989; International Working Group for Disease Monitoring and Forecasting, 1995).

Capture-recapture methodology is a general method to estimate the size of populations. An early example is estimating population specifics such as birth and death rates in an area near Calcutta, India (Sekar and Deming, 1949). The methodology has also been used regularly for Census purposes, for instance in the United Kingdom (Brown et al., 1999; ONS, 2012b) and in the US (Wolter, 1986a; Bell, 1993; Nirel and Glickman, 2009).

The capture-recapture methodology is well documented. The challenge, however, lies in the practical application of the methodology. For example, difficulties can arise in identifying which data sources best describe the population. Moreover, sometimes missing values are present in data on crucial variables, or assumptions of the method may be violated. Such difficulties can often not be avoided and solutions have to be found in order for the population size estimation to lead to correct outcomes.

This chapter is accepted as: Gerritse, S.C., Bakker, B.F.M., de Wolf, P-P., and van der Heijden, P.G.M. (2016). Undercoverage of the population register in the Netherlands, 2010. CBS publication. Acknowledgement of author contributions: theoretical development by BB, SG and PvdH. The analyses are carried out by SG. SG (main document) and PdW (variance estimation) wrote document, that was edited by BB and PvdH. The authors gratefully acknowledge support by Jan van der Laan (Statistics Netherlands) in linking the registers and other computational help, and Coen van Duin and Jeroen Pannekoek (Statistics Netherlands) and the CBS Advisory Committee on Census Under Coverage for their valuable comments.

Assessing the undercoverage of the Dutch Population Register asks for a definition of the Dutch population. Defining the rules according to which a person is, or is not, part of the population of a country has a lot of consequences, such as allocation of parliamentary seats in the EU and the attribution of funds depending on population size. Thus the definition of the population of a country is important statistical information and the Census is the primary framework to define a population (Lanzieri, Geneva, 30 September 3 October 2013). According to the United Nations Statistics Division (2008), we can define the population of a European country along the terms of usual residence:

"1.461. In general, "usual residence" is defined for census purposes as the place at which the person lives at the times of the census, and has been there for some time or intends to stay there for some time",

According to the European Union, Regulation (EU) No 1260/2013 of the European Parliament, usual residence is defined as

"The place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage".

An individual is considered a usual resident when they have lived in the Netherlands for a continuous period of 12 months before a Census reference time, or if they arrived in the 12 months before a Census reference time and intend to stay for at least a year. When these circumstances can not be established, "usual residence" means the place of registered residence (European Parliament, 2008). In accordance with the European Union regulations, in this manuscript we use the definition of residing for 12 months for usual residence. However, intention to stay is not registered and instead we define usual residence as residing more than 12 months continuously in the Netherlands.

This manuscript will document the steps needed for the application of capture-recapture methodology on the case of the undercoverage of the Dutch PR using usual residents. In doing so, we document the problems that arose in achieving this research goal, and how they were handled. This research poses as an example of how to deal with possible problems in the practical application of capture-recapture methodology in Official Statistics.

The Dutch Population Register (PR) used for the 2011 Census round was still the Gemeentelijke BasisAdministratie (GBA[1]) Under Dutch regulations, every individual residing in the Netherlands for longer than four months, or is planning to do so, should register in the PR. As such, the PR contains demographic information on the de jure population and differs from the de facto population, which is the actual number of individuals residing in the Netherlands regardless of registration. The coverage of the PR alone is not sufficient to provide a valid estimate of the 'de facto' population, in this manuscript defined by number of usual residents.

This incompleteness of the PR has more than one reason. First, within the European Union there is free movement and employment for individuals with an EU nationality. When an individual with an EU nationality resides in the Netherlands for longer than four months without having registered in the PR, they are not illegal residents, despite that they can be fined by Dutch law for not registering. Individuals that have not registered themselves may have forgotten to do so, or simply do not want to. These individuals are considered usual residents by the definition of the European Union but belong to the undercoverage of the PR. Second, the PR is also incomplete due to immigrants, coming from outside the European Union without a working or residence permit. These individuals then are illegally residing undocumented immigrants. These illegally residing undocumented immigrants are also considered usual residents, but are part of the undercoverage of the PR.

The registered population will also contain an overcoverage, and this may occur when registered individuals no longer reside in the Netherlands because of, for example, administrative delay of registering emigration and death. In the Netherlands, however, this is not as big a problem as the under coverage. (Bakker, 2009) estimated an over coverage of 31 thousand individuals, which is only 0.2 percent of the PR registered population.

To estimate the number of individuals missed by the PR, we linked three registers: the PR, an Employment Register (ER) and a Crime Suspects Register (CSR). In the ER jobs are documented. In the CSR suspects of all known crimes are registered. Unfortunately we can easily deduce residence duration for the PR only. For the ER usual residence can in part be deduced based on job lengths, assuming individuals residing in

[1] The GBA is currently replaced by the Basis Registratie Personen, (BRP). The BRP differs from the GBA because it also registers foreign individuals that have some sort of relationship with the Netherlands, and covers more individuals. This includes individuals with a Dutch nationality living abroad. Also, municipalities can register individuals that have not registered themselves, something which was not possible with the GBA. It is possible however is that the BRP, compared to the GBA, has a higher overcoverage of the Dutch population.

the Netherlands during the period that they hold a job. Consecutive and overlapping jobs were considered as one residence duration. However, for individuals that are unemployed between jobs, a decision had to be made on a length of unemployment that will be allowed between two jobs to still be considered as one continuous residence duration. The CSR has no information to deduce residence duration and for those individuals not linked to the PR and/or the ER we use missing data methodology.

During the linkage process it was found that 37% of the individuals in the CSR that did not link to both the PR as well as the ER, had incomplete linkage key information. It may be possible that these individuals are also in one or both of the other two registers but could not be linked because they had incomplete linkage key information. A part of these individuals could be erroneous captures. When individuals that are suspected of a crime are not registered in the PR, he police can not verify their information in the PR. Then, inaccurate information may occur, or even missing data. Therefore, when individuals have no crucial linkage key information, such as address and nationality, it is possible that these individuals do not belong to the population and, from our perspective, are erroneous captures. Hook and Regal (1995) and Gerritse et al. (2016) found that erroneous captures and linkage error can result in bias in the population size estimate resulting from capture-recapture analysis. In this paper, we propose different scenarios where, for the individuals in the CSR that cannot be linked and have missing data, the proportions of linkage error and erroneous captures are varied to assess the effect on the population size estimator.

Additionally, capture-recapture methodology relies heavily on a couple of assumptions Van der Heijden et al. (for example 2012), that can not be verified from the data. In this manuscript every assumption of capture-recapture methodology will be discussed and we will use the information and resources available to meet these assumptions as best as possible.

The manuscript is structured as follows. In section 2 the linkage of the data sources used will be discussed. In section 3 we will discuss residence duration. First, in section 3.1, the sensitivity analysis for the residence duration of the ER will be discussed, followed by, in section 3.2, the missing data methodology used to impute the residence duration of the CSR. In section 4 we shortly discuss capture-recapture methodology and its assumptions of capture-recapture analysis and the scenarios we use, and it discusses how the analysis is conducted. Section 5 gives the results from the capture-recapture estimation. The results will be discussed and concluded in section 6.

5.2 Linkage of data sources

We use three registers in this manuscript, the PR, the ER and the CSR. For the capture-recapture analysis only individuals that did not have a Dutch nationality were considered. Individuals with a Dutch nationality were considered Dutch residents and were excluded from the analysis. This also included individuals who had two nationalities, of which one of them was Dutch. We use ultimo September 2010 as the reference time point, or reference date.

For the purpose of our analyses the ER has been transformed into a register on individuals, where jobs were attributed to the individuals holding those jobs. The ER adds new cases of usual residents to the PR because it also registers individuals working in the Netherlands that have not registered themselves in the PR. Individuals with an EU nationality are free to work in other EU countries, but for salary and tax purposes employers will register these individuals in the ER.

The CSR documents individuals that are suspected of a known crime. In principle, this register can hold information on everyone in the Netherlands, including undocumented immigrants or other non PR registered individuals, because every individual residing in the Netherlands has a chance to become a suspect of a crime. Thus, it may provide cases that were not found in the PR but do belong to the population.

There is little information in the CSR and the ER on individuals with ages under 12, and over 65, because individuals under 12 can not be registered in the CSR and the ER only registers between 15 and 65: As such it was decided that the population specified in this paper will consist only of the population aged between 15 to 65.

To link the three registers we used two types of linkage, deterministic and probabilistic, both of which are considered a type of exact matching (Herzog et al., 2007). For a pair to be a link under deterministic linkage, the two records have to agree exactly on each element in a set of identifiers. One example is when two records match on name, address, date of birth and city of birth, or on a Personal Identification Number. Fellegi and Sunter (1969) formalized probabilistic linkage where all records in one register are paired to all records of a second register. Based on their agreement on a set of identifiers, a weight is given to each pair. A predefined cut-off determines whether pairs are links, non-links or possible links (Herzog et al., 2007). To reduce the number of possible pairs for computational efficiency we can block the data on one, or more variables.

At statistics Netherlands all registers are linked deterministically to the PR via a PIN and a linkage key. The PR is the backbone of Statistics Netherlands, such that all registers and surveys are linked to the PR. Linking other registers, of which also the ER and the CSR, to the PR via a Dutch PIN, enables about 96 to 98 percent of the cases to be linked to the PR. When a case has no PIN, the registers are linked on postal code, house number, date of birth and sex. Then 93 to 95 percent of individuals can be linked to the PR (Arts et al., 2000). Thus it seems that when cases have complete linkage keys, linkage to the PR can be done in more than 90 percent of the cases. However, there are cases with incomplete linkage key information, such that they can not be linked. These cases are often without a Dutch nationality and are in the subpopulation of interest for this study. To improve upon the deterministic linkage we also used probabilistic linkage.

For the individuals in the PR and ER we used a combination of date of birth, sex, postal code, house number and suffix as linkage key to probabilistically link the remaining ER units to the whole of the PR. This resulted in an overall improvement of 0.1% of the remaining ER-units that can be linked. The same linkage key is used to probabilistically link the remaining units in the CSR to the whole of the PR, which led to an overall improvement of 3.3% of extra individuals linked. For both the linkage of the ER to the PR and the CSR to the PR we blocked first on postal code or date of birth to reduce the number of possible pairs. To probabilistically link the whole of the ER to the whole of the CSR we used a combination of birth date, sex, city of residence, country of residence, street name and house number as linkage key. For the probabilistic linkage of the ER to the CSR, the data was blocked on either city of residence of birth date had to be equal to at least day or month of birth and led to an overall improvement of 9.9% of extra individuals linked.

Figure 5.1 shows the linkage of the PR to the ER, the PR to the CSR and the ER to the CSR, and more specifically the percentage of each register linked either deterministically (in yellow) or probabilistically (orange). It was found that 37.7% of the individuals in the CSR that could not link to the PR or the ER had incomplete linkage key information. Thus these individuals were unable to be linked.

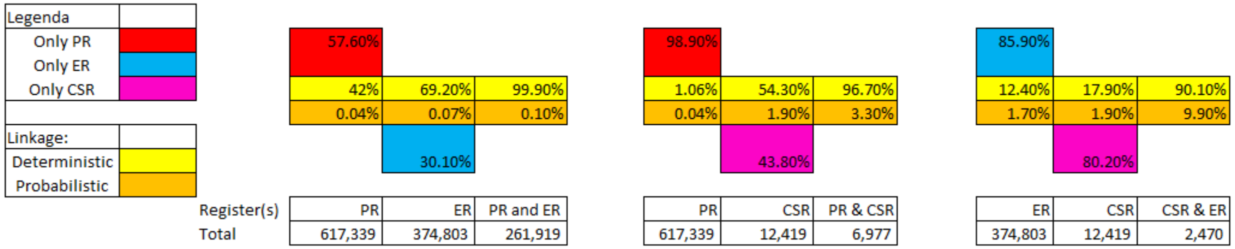

FIGURE 5.1
Linkage of the PR to the ER, the PR to the CSR and the ER to the CSR. Of the first figure, the first column shows all the individuals in the PR, where the second column shows all the individuals in the ER. The overlapping parts in yellow represent the percentage of individuals that have been linked deterministically, and the parts in orange represent the percentage of individuals that have been linked probabilistically. The last column shows the linked part between the PR and the ER, where it can be seen that probabilistic linkage led to an improvement of 0.1%. The other two figures show the same kind of information for respectively the PR linked to the CSR, and the ER linked to the CSR.

We use four covariates for capture-recapture analysis: nationality group, age, sex and usual residence. Nationality group has 7 categories: (1) EU15 (excl. Netherlands) (2) Polish (3) Other EU (4) Other western (5) Turkish, Moroccan, Surinam (6) Iraqi, Iranian, Afghan, asylum seeker countries Africa (7) Other Balkan, former Soviet Union, other Asian, Latin American and other nationalities. The countries are clustered according to likely migration motives, migration legislation, regulations of the PR and size. For age, we use four levels: (1) 15-24 (2) 25-34 (3) 35-49 and (4) 50-65.

TABLE 5.1
Observed values for the three registers.

PR	ER	CSR		Total
		Yes	No	
Yes	Yes	2,115	259,804	261,919
	No	4,862	350,551	355,413
No	Yes	355	112,529	112,884
	No	5,087	0	5,087
	Total	12,419	722,884	735,303

Table 5.1 shows the counts for the individuals in the three linked registers ignoring the distribution over the four covariates, as was also used in Gerritse et al. (2015a). Table 2 shows that the individuals are not evenly distributed over the three registers, which may lead to small cell counts in the capture-recapture analysis when the four covariates are taken into account. This may complicate the population size estimation. It was found in previous research, which also used this data set, that even the smallest changes in the data can result in different population size estimates (Gerritse et al., 2015a). For that purpose, we use different scenarios to account for the biggest variability in this manuscript, and assess the impact on the population size estimation.

5.3 Methods for deriving residence durations

Neither register has a direct measure of residence duration. For the PR a measure of residence duration can be deduced. The PR has a registration date, the date at which either a person was born into the Netherlands or when immigrants registered themselves in the PR as a Dutch resident.

Then the reference date of ultimo September 2010 minus the registration date can be used as a residence duration. Additionally, when a PR registration was consecutive or overlapping with a job, such as when a job was registered in the ER before the PR registration, the combined length of PR and ER registration was also considered a residence duration. The following sections will document the steps taken for the usual residence measure of the ER and the CSR.

5.3.1 Residence duration ER

In the ER, amongst others, the start date and the end date of the job are registered. Assuming that individuals that have a job in the Netherlands, also reside in the Netherlands during the time they hold that job, we can use the information on successive jobs previous to the reference date as a residence duration. For individuals that held more than one job that were consecutive or overlapping we perceive these jobs as one period of work and therefore residence. A total of 77% of the 374,803 individuals in the ER had one period of residence.

When there was a period of unemployment between two jobs, the question is whether this affects an individual's usual residents status. For the individuals in the ER that did not link to the PR, and thus did not have a residence duration from the PR, a decision had to be made on the length of unemployment between jobs that would be allowed, wherein the individual still perceived as residing in the Netherlands. We investigated seven scenarios. A total of 1, 8, 15, 22, 31, 62 and 93 days were allowed between two jobs for an individual. In the first column of Table 5.2 are the observed values of all individuals ER by nationality groups. The other columns show for each of the scenarios the observed number of individuals in the ER but not in the PR that reside in the Netherlands for longer than a year. As can be seen, for an individual with a EU15 nationality 7,781 individuals are considered residing in the Netherlands for longer than a year based on their job length when we accept an unemployment of 8 days between two jobs. If we increase this to 15 days the number of individuals residing in the Netherlands increases to 7,915, up to 10,861 when we consider 93 days. The different scenarios give different numbers of usual residents.

TABLE 5.2
Scenarios for deducing residence duration from joblengths in the ER, per nationality group. The first column shows the observed counts of all the individuals in the ER that are not in the PR. The other columns show the count of individuals that would be considered usual residents under the specified scenario

Nationality	Total	1 day	8 days	15 days	22 days	31 days	62 days	93 days
EU15	18,727	7,548	7,781	7,915	8,093	8,862	9,982	10,861
Polish	80,738	14,711	15,677	16,301	16,904	20,315	25,520	29,399
Other EU	10,765	2,201	2,267	2,331	2,369	2,617	3,095	3,420
Other West	509	216	218	220	221	222	231	242
Turkey, etc.	647	341	349	354	363	380	425	457
Iraq, etc.	274	142	145	149	155	166	179	194
Balkan, etc. Other.	1,378	530	539	543	553	567	615	671
Total	113,038	25,631	27,034	27,871	28,717	33,193	40,126	45,325

It is difficult to choose which scenario will be more realistic. As such a choice has to be made on the few indications that can be found in Table 5.2. We have chosen for the scenario of 31 days. A choice could be made for a smaller scenario, but it seemed more realistic to allow for one month, given that jobs often start at the beginning of a month and end at the end of the month. Thus allowing at least a month is plausible. Also, it can be seen from Table 5.2 that the biggest, absolute, increase in usual residents is between 31 and 62 days, which indicated a turning point between these scenarios where more individuals may decide to leave the country for a longer period. The biggest group of non-Dutch individuals is of European nationality, and the probability that they return to their home country after 31 days increases. Additionally, it is slightly more conceivable that when individuals are unemployed for up to 31 days that they are still in the Netherlands, compared to when more months are allowed. A total of 113,038[2] individuals in the ER are not in the PR, of which we consider 33,135 individuals are usual residents. The remaining 79,903 individuals in the ER that did not link to the PR are non usual residents. Note that for the individuals that are registered in the ER and that do link to the PR, we have two residence durations, one from the PR registration and one deduced from job lengths in the ER. Of these two measures of residence duration, we chose the one that is the longest, either the PR residence duration or the ER job lengths duration.

5.3.2 Residence duration CSR

For the CSR it is not possible to define a residence duration based on the variables in the register itself. Moreover, the CSR is an event based register in which crimes are recorded on one particular date. Most of the suspects are first offenders, which means that there is only one date recorded for most individuals and a residence duration can not be deduced from this one date. For the records that link to the PR and/or the ER, we use the residence duration from these sources. When usual residence was available from both the PR and ER, the longest residence duration of either the PR or the ER was chosen. However, for the remaining records in the CSR, residence duration is missing, and we have to impute it.

In earlier research it was found that the missing residence duration values for the CSR will probably resemble the usual residence values from the ER more than the usual residence values from the PR (Gerritse et al., 2015a). The CSR records will resemble the ER records rather

[2] After probabilistic linkage of the ER and PR to the CSR it was found that 154 ER registered individuals, and 7 PR registered individuals, were erroneously seen as CSR registered individuals only.

than the PR records, because the individuals that register themselves in the PR have the intention to stay for a longer period in the PR. This can not be assumed for the individuals in either the ER or the CSR that do not link to the PR, given there is a reason why these individuals have not registered themselves. As such we need an imputation method that imputes missing data with this very specific subset of data in the ER that it resembles most. It was found that for our specific missing data problem, multiple imputation by Predictive Mean Matching (PMM) will most likely handle the missing data best, given that PMM allows the user to ignore individuals in the PR (Gerritse et al., 2015a).

Predictive Mean Matching is an example of a hot deck, nearest neighbour multiple imputation method. For each missing entry, PMM samples a small set of candidate donors from the observed complete cases (Buuren, 2012). The predicted values are estimated based on the information in the predictor variables and their "closeness" is estimated by their absolute difference. From this small set of cases, one is randomly chosen to replace the missing value. The assumption here is that the missing value follows the same distribution as the observed values chosen in the small set (Buuren, 2012).

This means that we use all 113,038 records from the ER and CSR that did not link to the PR to impute the missing usual residence value for the individuals in the CSR only. These donors will be used to impute the missing residency variable in the CSR units that do not link to either of the other sources. PMM has been repeated ten times, to account for multiple imputation variability. The PMM multiple imputation has been done via the package mice in R (Buuren, 2012)

5.4 Method for estimating the number of unregistered individuals

Capture-recapture methodology is an often used methodology for population size estimation. The simplest form of capture-recapture methodology consists of linking two sources of data, such as samples, lists or registers. Assume we have two registers. Record linkage finds cases in both registers that are from the same individual by using identifying variables, such as a PIN or a combination of linkage key variables like name, address, etc. The result is a count of individuals in either register alone or in the overlap of both registers. Additionally there is a zero count of individuals that do belong to the population but were not observed in either register.

Assume we have two registers, register 1 and register 2. Let i = (0, 1) respectively denote not included in register 1 and included in register 1. Also, let j = (0, 1) respectively denote not included in register 2 and included in register 2. Let m_{ij} denote the expected values for registers 1 and 2, and n_{ij} denote the observed counts. Then odds ratios can be used assuming independence between register 1 and register 2, such that $m_{11}m_{00}/m_{01}m_{10} = 1$. However m_{00} is a structural zero and is the value we are interested in. Assuming independence, this odds ratio can be rewritten to get maximum likelihood estimate

$$\hat{m}_{00} = \frac{\hat{m}_{01}\hat{m}_{10}}{\hat{m}_{11}} = \frac{n_{01}n_{10}}{n_{11}}. \tag{5.1}$$

This odds ratio is for two registers only, but can be easily extended from 5.1 to the three register case. Let m_{ijk} be expected values for registers 1, 2 and 3. Let variable C denote inclusion in register 3, such that k = (0, 1) respectively denotes not included in register 3 and included in register 3. Then odds ratios are used assuming the three factor interaction to be absent, so that $m_{110}m_{000}/m_{101}m_{011} = m_{111}m_{001}/m_{100}m_{010}$. Expected value m_{000} is a structural zero and can be estimated by

$$\hat{m}_{000} = \frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}}. \tag{5.2}$$

Equations 5.1 and 5.2 assume saturated loglinear models and can be readily implemented. More parsimonious models, compared to the saturated model, are possible and may fit the data better. Such models will be used in capture-recapture estimation later in this manuscript.

The capture-recapture analysis has been conducted via Generalized Linear Modeling (GLM) in R. The function STEP in R enables the researcher to select the best fitting loglinear model. By default STEP selects models based on the AIC. However since our sample size is quite large, the AIC will lead to models that are unnecessarily complicated. Therefore we used the BIC, since the BIC has a larger penalty for sample size. The details on how the variance for confidence interval testing was estimated can be found in the Appendix.

A summary has been made of previous research on the past number of unregistered individuals in the Netherlands in Gerritse et al. (2015a). We present here only the conclusion of this summation, as the arguments are presented in that manuscript. We expect that the under coverage of the PR will lie within a range of 175 to 225 thousand individuals. This range has been deduced from research by Hoogteijling (2002), Bakker (2009) and Van der Heijden et al. (2012), and migration flows and asylum requests since these research articles. We acknowledge that this range is based on a couple of assumptions that may not be met, and we use this range with caution as an indication where the undercoverage may lie. Thus possible estimates of the under coverage of the PR that falls outside this range is not necessarily false. However, when our estimates lie far from the lower or upper boundary of this range, the estimate does become rather implausible.

5.4.1 Implied coverage

Capture-recapture relies heavily on a couple of assumptions (Van der Heijden et al., 2012). Violation of these assumptions could lead to biased estimates, as was found in Brown et al. (2006); Boden (2014); Gerritse et al. (2015b, 2016). These researchers found that sometimes violated assumptions have little effect on the population size estimate, whereas sometimes it had a large impact on the population size estimate. The effect a violated assumption has on the population size estimator has been found to be a direct result of the implied coverage of the main register (Gerritse et al., 2016). Implied coverage describes the observed coverage of register 1, given register 2. As such, it describes the number of new cases added by register 2, compared to the already known cases in register 1.

From equation (5.1) we can estimate conditional probabilities $\hat{p}(0|1) = n_{01}/n_{+1}$ and, $\hat{p}(1|1) = n_{11}/n_{+1}$. Thus $\hat{p}(0|1)$ is the estimated probability of new cases from register 2, among all cases in register 2, and $\hat{p}(1|1)$ is the estimated probability of already known cases from register 1, among all the cases from register 2. Then equation (5.1) changes to

$$\hat{m}_{00} = \frac{\hat{p}_{0|1} n_{10}}{\hat{p}_{1|1}}. \tag{5.3}$$

When $\hat{p}(0|1)$ is relatively small compared to $\hat{p}(1|1)$, the effect of the added new cases of register 2 on $\hat{m}_{00}$ will be small and the population size estimator is robust to violations of the assumptions. However, when $\hat{p}(0|1)$ is relatively large compared to $\hat{p}(1|1)$, the effect of the added new cases of register 2 on $\hat{m}_{00}$ will be large as well and the population size estimator is not robust to possible violations of the assumptions.

The use of a third register and multiple covariates make it more complex to investigate the effect of the implied coverage on the population size estimator. From Table 5.1 we cautiously conclude that the implied coverage of the PR given the ER, and the implied coverage of the PR given the CSR seems relatively low. A low implied coverage will result in a higher number of individuals in the numerator of equations (5.1) and (5.2), compared to the denominator, and will result in an unstable estimate. To avoid any effect implied coverage has on the population size estimator when assumptions are violated, extra care has been taken to make sure the assumptions were met.

5.5 Implied coverage for three registers

Implied coverage can be extended to the three register case. Assume that register 1 covers most of the population, then register 2 and then register 3. For our purposes it suffices here to discuss coverage of registers 1 and 2 implied by the third register. For this purpose Equation (5.2) is complicated, as n_{111}, the number of cases seen in all three registers, is in the numerator. We focus on the observed counts n_{101}, n_{011} and n_{001}, i.e. the number of individuals seen only in register 3, i.e. n_{001}, compared to the individuals seen in register 3 and only in register 1, i.e. n_{101}, and compared to the individuals only seen in register 3 and only in register 2, i.e. n_{011}. We focus first on n_{001} and n_{101}. Notice that these counts refer to individuals missed by register 2. Therefore we will speak of coverage conditional on being missed by register 2. Now, similar to the discussion of implied coverage in a two-way table above, if n_{001} is large in comparison to n101, the conditional coverage of register 1 implied by register 3 is low, where conditional refers to being missed by register 2. Thus the estimator in equation (5.2) becomes unstable when the number n_{001} becomes unstable. Similarly, for n_{001} and n_{011}, if n_{001} is large in comparison to n_{011}, the conditional coverage of register 2 is low and the estimator in equation (5.1) becomes unstable when the number n_{001} becomes unstable (where conditional refers to being missed by register 1. Anyhow, for all practical purposes it will be clear that when n_{001} is large, the estimator defined in (5.2) will become unstable.

Table 5.1 reveals that the implied coverage of the PR, given the ER is relatively high. The implied coverage of the PR, given the CSR however is rather low. The implied coverage of the ER, given the CSR is also low. Summarizing, the conditional coverage implied by the CSR is low. This indicates that the population size estimator is not robust to possible violations of the assumptions.

5.5.1 Assumptions

Capture-recapture analysis relies heavily on a couple of assumptions of which can not be verified from the data whether they are met. Above we stated that implied coverage seems low and thus the estimator may not be robust to possible violations of the assumptions. As such we discuss below the extra steps taken to make sure the assumptions are met as best as possible.

Closed Population

The first assumption is that the population is closed. For the PR and the ER this assumption can be easily met. The PR and ER contain information on individuals over a period of time, such that one timepoint can be chosen to keep a closed population. In our case the timepoint chosen is the reference date of ultimo September 2010. The CSR however is a register containing solely reports on suspects of all known crimes, and one specific day cannot be chosen since on that day only a limited number of of reports could have been filed.

Thus a specific time period has to be chosen in which observations are taken from the CSR. This makes it difficult to assume a closed population. For the CSR a time period of half a year, the second half of 2010, has been chosen. We have chosen for half a year because then the number of individuals in the CSR is still relatively large, but also this time period is short enough so that a violation of the closed population assumption will be relatively minor. The second half of 2010 has been chosen because then the reference point used for the PR and ER lies in the middle of the six months considered.

Independence and homogeneous inclusion probabilities

Under independence one assumes that the probability to be included in the first register is independent on the probability to be included in the second register. This is a rather strict assumption. There are two ways to relax the independence assumption. One way is in using multiple sources. In using three registers the strict independence assumption is relaxed into the assumption that the three way interaction is zero. Additionally, in adding covariates heterogeneity due to these covariates can be removed (compare, International Working Group for Disease Monitoring and Forecasting, 1995). In this manuscript we use three registers and four covariates, i.e. age, sex, nationality group and residence duration, to relax the strict independence assumption and account for possible heterogeneity in these covariates.

It has to be noted that the covariate usual residence is operationalised differently per register. Additionally, when cases were in the overlap of two registers, the longest usual residence value was chosen, and as such the operationalisation of usual residence in the overlap of registers is different to the operationalisation of the individual registers as well. Then dependence may well be introduced in usual residence considering the registers. In using usual residence as a covariate in the loglinear model in estimation, this dependence can be accounted for.

Perfect linkage

In this manuscript two methods for linkage have been used to increase the probability that we linked all cases that had to be linked. At Statistics Netherlands every register is linked deterministically to the PR, such that the PR and the ER, and the PR and the CSR, were already linked deterministically. However, due to errors in the variables used for linkage of the CSR and ER it was difficult to deterministically link the CSR to the ER. To improve the deterministic linkage, probabilistic linkage was used (see section 2 for more details).

No erroneous captures

The last assumption considered here is that the registers only contain information on the specified population, and thus do not contain any erroneous captures. When registering in the PR an individual has the intention to stay for a longer period, and thus they are assumed to be usual residents. Then for the PR we assume that there will be no erroneous captures. Individuals with an address in Belgium or Germany are removed from the analysis, as we assume that individuals from our neighbouring countries will probably still live in Belgium and Germany and only cross the border for work, school or possibly crime related activities. Additionally individuals are removed that were caught by the border police and thus did not enter the Netherlands at all.

Concluding remarks assumptions

It was found that 37 percent of the individuals in the CSR that did not link to the PR and the ER had missing or incomplete linkage key information. There is a chance that these individuals do belong to the population and had to be linked to the PR and the ER. There is also the possibility that these individuals have missing information for the reason that they do not belong to the population and thus are erroneous captures. Unfortunately, we cannot deduce from the data to which extent the assumption that there are no erroneous captures is violated. Additionally, we found that implied coverage is low due to a large number of individuals in the CSR that cannot be linked to the PR and the ER,, and as a result the population size estimator will not be robust to possible violations of the assumptions. For that reason, we will investigate via scenarios which percentage of linkage errors and erroneous captures seem most realistic.

For the individuals in the CSR that did not link to the PR and the ER there is still the probability of duplicate cases. Due to the incomplete linkage key information we cannot establish the number of duplicates. We can however treat such cases as erroneous captures, and investigate the effect of removing these individuals from the analysis.

5.5.2 Scenarios

Main scenarios

Because implied coverage for the three registers is low, we use different scenarios to assess the effect of the presence of linkage error and erroneous captures in the 37% of individuals in the CSR that had incomplete linkage key information on the population size estimate. A baseline scenario will be set up where we assume no linkage error and no erroneous captures, and consider all 37% to belong to the population and to have only been registered in the CSR. For scenario 1 and 2 we will divide up the 37 percent of individuals in the CSR without linkage key information as either linkage error or erroneous captures.

The probability is higher that the individuals with incomplete linkage key information do not belong to the population and are more likely to be in the Netherlands as a tourist or for criminally related purposes. The police uses the PR to identify suspects. When the police cannot find suspects in the PR, they will denote whatever information is available on the suspect. It is plausible that the police cannot denote information for individuals that do not belong to the population. When these individuals have no Dutch residence it will be a hard task to procure such information, resulting in missing information.

Methodologically this argumentation can be supported by the recent investigations of Zhang (2015) and Zhang and Dunne (2015), who researched capture-recapture methodology in the presence of erroneous captures. A trimmed dual system estimation method was set up. Erroneous captures are identified by linking the register to a post enumeration survey (PES). By assuming that the PES does not have erroneous captures, the size of the over coverage due to erroneous captures can be estimated. The next step is that records are to be excluded from capture-recapture analysis. The trimmed dual system estimator reduces the bias caused by erroneous captures by deleting them from the analysis. However, identifying erroneous captures is a hard task and the researchers advise against randomly deleting cases. Knowing that the 37 percent have no linkage key information, removing most of these individuals is a first step in deleting erroneous captures.

Therefore, we assume that for scenario 1 and 2 the majority of the 37 percent are erroneous captures. In scenario 1 we will consider a random selection of 75% of the individuals without linkage key variables as erroneous captures and remove them from the analysis, and the other 25 percent will be considered as linkage errors and these individuals from the CSR that did not link to the PR and ER will be linked to the PR and ER. Scenario 2 will investigate the effect of removing all individuals without the linkage key variables, the total of 37 percent, as though they are all erroneous captures.

Subscenarios

Additionally, there is a possibility that of the 63 percent of individuals in the CSR that did not link to the PR and ER, and do have complete linkage key information, there still are some linkage errors or erroneous captures. The CSR might contain possible administrative errors that have to be dealt with, even though these errors will be small. As such, scenarios 1 and 2 will have four possible outcomes, as shown in Table 5.3. Subscenario *a* from Table 5.3 (scenarios 1a and 2a) will have no linkage error or erroneous captures taken from the 63% individuals with complete linkage key variables. This will provide a baseline estimate for the scenario. In subscenario *b* (scenarios 1b and 2b from Table 5.3) an extra 5 percent of erroneous captures will be removed from the individuals with complete linkage key information and in subscenario *c* (scenario 1c and 2c from Table 5.3) an extra 5 percent of linkage errors will be taken from the individuals with complete linkage key information. In subscenario *d* (scenario 1d and 2d from Table 5.3) both an extra 5 percent of erroneous captures and linkage errors will be taken from the individuals with linkage key information.

For the individuals under subscenarios c and d that are considered linkage error we have to decide to which register the individuals in the CSR would link. We have stated before that we think that the individuals in the CSR resemble those in the ER most that did not link to the PR. As such, when individuals in the CSR were linkage error, the probability will be higher that they should have been linked to the ER, rather than to the PR. Due to possible errors, some will however have been linked to the PR. It has been chosen that 80 percent will be linked to the ER. The remaining 20 percent will be linked to either the PR or the PR and ER in accordance with the distribution of how the observed data are distributed over the PR or the PR and ER. Given that 60 percent of all observed PR values do not link to the ER and 40 percent do, 60 percent will be linked to the PR and 40 percent will be linked to the PR and the ER.

Note that the individuals will be randomly assigned to link to either the PR, ER or both registers. This will have an effect on the population size estimator. However, this effect will be considerably smaller than randomly assigning all of the 37% of individuals as either linkage error or erroneous captures because the number of observations considered are smaller.

5.5.3 Analysis

The analysis has been conducted as follows. The dataset used contained all the individuals in the PR, ER and the CSR that did not have a Dutch nationality. We also removed individuals that had a German or Belgium registered place of residence and individuals apprehended by the border police at Schiphol. Usual residence for the PR and ER have been deduced as described in section 3. For the individuals in the CSR that did not link to the PR and ER, usual residence was still missing.

Table 5.1 shows that we have 5,087 individuals in the CSR that did not link to the PR and ER. Of these 5,087 individuals, 1,917 individuals (37%) have an incomplete linkage key. Scenario 1 is investigated as follows. From the 1,917 individuals 75% will be considered erroneous captures and 25% will be considered linkage error. First, we remove the 75% of erroneous captures, which entails 1,438 of the 1,917 individuals from the CSR that did not have complete linkage key variables. Then 25% of the individuals in the CSR without complete linkage key variables are considered linkage error, which entailed 479 individuals. As was discussed in section 4.4, we assume 80 percent of the individuals to have linked to the ER, 12 percent to the PR and 8 percent to the intersection of the PR and ER, such that 383 individuals will be linked to the ER, 57 to the PR and 38 to the PR and ER.

Every scenario has a subscenario, where we consider also 5% linkage error and/or erroneous captures from the 3,170 individuals, or 63 percent, that did have complete linkage key values. Assume we took both, and we operate in scenario 1d. Then first, from the $5,087-1,917 = 3,170$ we delete 5% which are considered erroneous captures, such that 158 individuals are removed from the data. From the 3,170 - 158 = 3,012 individuals remaining, another sample of 5% is considered linkage error. Note that this is 5% of the 3,170 and not 5% of the 3,012 individuals remaining after having already taken 5% erroneous captures. Of these

158 individuals, we link 80 percent, or 126 individuals, to the ER, 12 percent to the PR (19 individuals) and 8 percent to the PR and the ER (13 individuals).

After having dealt with the simulated linkage error and erroneous captures we conduct the multiple imputation using Predictive Mean Matching. We use the package MICE to impute the missing values on the variable usual residence for the CSR using every variable (except the PR) in the imputation model. Mice has been programmed to impute ten times with only one iteration, given that there was only one variable that had missing values.

What follows is carried out for $M = 10$ imputations and $N = 7$ nationality groups. First, a loglinear model has been estimated using main effects only. This model is used as a starting point to use the function STEP which selects the best fitting loglinear model on the data. The selection has been done using the BIC. The best fitting loglinear model is used to estimate the size of the population missed. The estimates are stored, averaged over the $M = 10$ imputations, and reported in this manuscript. The estimate on the total missed portion of the population is found by summing the estimates of all 7 nationality groups. For the two estimates for which a sampling analysis will be done, the analysis also includes a bootstrap with 10,000 iterations to estimate the variance of the resulting estimate. However, we use multiple imputations and for some estimates also multiple samples. We also need to take into account the variance due to both multiple imputation and sampling. The variance for multiple imputations has been coined by (Little and Rubin, 2002), and an extension from this variance to incorporate variance from multiple samples can be found in the appendix.

5.6 Results

The results can be found in Table 5.3. The maximum likelihood estimates of the missed portion of the population that is estimated by capture-recapture methodology vary considerably over the scenarios. When we consider all 37% of individuals with an incomplete linkage key as belonging to the population that were only registered in the CSR, thus the baseline scenario, the three registers miss a total of 249 thousand individuals. However, when we consider 75% of individuals that have an incomplete linkage key as erroneous captures and the remaining 25% as linkage error the estimate drops to 66 thousand individuals. Thus when we remove 1,438 of the 1,917 individuals as erroneous captures and link the remaining 479 individuals to the other two registers, the resulting population size estimate is rather different than when we consider these 1,917 individuals to belong to the population that are only registered in the CSR. When we consider the full 37% of individuals with an incomplete linkage key as erroneous capture the estimate becomes 151 thousand.

TABLE 5.3
Overview of the scenarios and the resulting maximum likelihood estimates of the missed portion of the population.

	37% with incomplete linkage key variables		63% with complete linkage key variables		
Scenario	Err. Capt.	Link. error	Err. Capt.	Link. error	Mle
					x1000
Baseline	0%	0%	0%	0%	249
1a	75%	25%	0%	0%	66
1b	75%	25%	5%	0%	66
1c	75%	25%	0%	5%	54
1d	75%	25%	5%	5%	56
2a	100%	0%	0%	0%	151
2b	100%	0%	5%	0%	151
2c	100%	0%	0%	5%	91
2d	100%	0%	5%	5%	92

Interestingly, linking an additional 5 percent of the 3,170 individuals with complete linkage key information to the PR and/or ER decreases the population size estimate considerably, whereas removing an additional 5% of erroneous captures from the 3,170 individuals with complete linkage key information does not. From both scenarios it is obvious that linkage error has a bigger effect on the population size estimation than erroneous captures do.

From the scenarios, not considering the baseline scenario, we see that the lowest estimate was 54 thousand and the highest estimate was 151 thousand individuals missed by all three registers. We found that there are 33 thousand usual residents in the ER that were not registered in the PR and thus are part of the undercoverage of the PR. Of the 5 thousand individuals in the CSR that did not link to the PR and ER, approximately 1 thousand individuals are residing longer than a year and have to be considered as part of the under coverage of the PR. Thus, there are 34 thousand usual residents in the undercoverage of the PR that are registered in the ER and CSR and have to be added to the estimates from the analysis to know the undercoverage of the PR.

This means that for the lowest estimate there are 88 thousand individuals in the under coverage of the PR, with a confidence interval of 57 to 151 thousand. For the highest estimate we have 185 thousand individuals not covered by the PR, with a confidence interval of 149 to 222 thousand individuals. Given that there are 16,638 thousand individuals registered in the PR, the undercoverage of the PR of 88 to 185 thousand usual residents means that we have an under coverage of the PR of only .5 to 1.1 percent.

5.7 Discussion

We are interested in estimating the number of usual residents via capture-recapture analysis to estimate the under count of the PR. We documented in this manuscript the steps taken to overcome the practical challenges that arose during the analyses. As such this manuscript is an example of how the well known capture-recapture methodology can be used in a practical applications such as estimating the undercoverage of a register.

It has to be noted that the data that was used in this manuscript was of individuals aged 15 - 65 years. This was due to restrictions in the registers. As such, the population size estimates and the resulting under coverage of the PR that are described in this manuscript consider only a population of 15 to 65 years of age. To estimate children up until 15 years of age and the elderly over 65 years, other information is needed.

In section 4.2. we found that implied coverages of the PR en the ER are low, which was caused by of the large number of individuals in the CSR that could not be linked. Because of the low implied coverage, the population size estimator will not be robust to possible violations of the assumptions and different scenarios were used to deal with the 37% of individuals in the CSR that had incomplete linkage keys. In the scenarios we randomly assigned cases as either erroneous captures or linkage error because there is no information to know which case belongs to the population and which does not. Unfortunately this random sampling does have an impact on the population size estimate and the variance of the estimate. The confidence interval for our highest estimate becomes as high as 185 thousand. This means that we have an under coverage of the population aged 15 to 65 of maximally 1.1 percent. Even though the confidence interval is high, the under coverage remains rather small.

In Gerritse et al. (2015a) we determined a range of possible outcomes, where we expect the population size estimate would lie within. The range of maximum likelihood estimates given in this manuscript includes the lower part of this range of possible outcomes of 175 to 225 thousand individuals. This range of possible outcomes has been created on outcomes of previous research on foreigners and their residence duration. A couple of assumptions had to be made, given that these previous research did not always had the information needed, such as residence duration (Hoogteijling, 2002), or only had information on part of the data (Van der Heijden et al., 2012). Thus, even though it is interesting to note that the ranges overlap, implications from this are not straightforward.

It is interesting that the range of possible outcomes of previous research overlaps with the current estimates of 88 to 185 thousand individuals. This may indicate that there is higher possibility that the actual number of usual residents missed by the PR lies close to the upper end of the range of 88 to 185 thousand individuals. If the actual number of under coverage is in the upper end of our range, there is only about a 1 percent undercoverage of the PR of the population of Dutch usual residents aged 15 to 65.

Extra precautions have been taken to make sure all assumptions have been met. Also, our estimates overlap with a range of possible outcomes by former research. However, it has to be acknowledged that our population size estimates may be biased still, due to unknown violations.

In this manuscript we have given an overview of what was deemed best for the data used at Statistics Netherlands to estimate the under coverage of the PR. The results and their implications do not only apply to the data used in this manuscript, it can be used as a caution for other research as well, especially the more similar their data is to this research. It has to be kept in mind that the capture-recapture methodology can be very sensitive to small changes in the data, but also in the estimation process. It is a useful method yet has strict assumptions which have to be taken into account, but it also becomes more complex when missing data is introduced. In this manuscript we have given our view on a possibility to work with this.

Appendix - variance estimation

For the maximum likelihood estimates the analysis also included a bootstrap with 10,000 iterations to estimate the variance of the resulting estimate. This bootstrapped variance is used as the within variance to estimate the total variance for multiple imputation, which is the left hand side of equation (5.5). In a last loop the stored estimates and variance are used to estimate the variance via (5.5). This variance is used to derive a 95% confidence interval on the two chosen estimates.

For the estimate of 151 thousand individuals missed (scenario 2a), all individuals with 37% of incomplete linkage key are considered erroneous captures. Thus there is no random assignment of erroneous captures or linkage error and we do not take extra samples to account for variability due to random assignment. For that scenario the variance needed for the confidence interval is estimated via the variance estimation formula of Little and Rubin (2002) that takes into account variance due to multiple imputations.

However, for the estimate of 54 thousand individuals (scenario 1c) there is random assignment of both the linkage error and erroneous captures of the 37% of individuals that have incomplete linkage keys, as well as an additional 5% of linkage error for the individuals in the CSR that did have a complete linkage key. In this manuscript we have multiple sources that affect the variance of the estimate. To assess the variance from multiple imputation, Little and Rubin (2002) have formulated a total variance that combines the within and between imputation variance. Thus the variance for multiple imputation can be accounted for. However, because we also take different samples of erroneous captures and linkage errors we want to account for that variance as well. We formulate here how the total variance formula from Little and Rubin (2002) can be extended to a third source of variance.

Consider the situation where we have taken a sample d of records, to simulate removing erroneous captures. For this sample we perform $i = 1, \dots, M$ imputations. Per imputation we can calculate the capture-recapture estimator $\hat{\theta}_d^i$. An estimate of the variance of $\hat{\theta}_d^i$, conditional on the taken sample, is obtained by applying a parametric bootstrap resampling procedure. We will denote this variance by

$$W_d^i = \widehat{\text{Var}}\left(\hat{\theta}_d^i\right).$$

The final estimator for sample d is averaged over all imputations:

$$\hat{\theta}_d = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_d^i.$$

Applying the formula for the total variance of $\hat{\theta}_d$ in case of multiple imputation as given e.g. in Little and Rubin (2002) yields

$$\widehat{\text{Var}}(\hat{\theta}_d) = \frac{1}{M}\sum_{i=1}^{M} W_d^i + \left(1 + \frac{1}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_d^i - \hat{\theta}_d\right)^2.$$

We thus have an estimator of the variance for the estimate of θ based on a single sample, with M imputations. Denote this variance by

$$W_d = \widehat{\text{Var}}(\hat{\theta}_d).$$

To be able to estimate the variance due to sampling records to simulate erroneous captures, we will replicate the procedure D times. That is, we will perform the sampling D times and for each sample $d = 1, \ldots, D$ we will calculate $\hat{\theta}_d$ and its variance estimate W_d. The total variance we will then estimate by applying the general total variance formula

$$\text{Var}\, X = \mathbb{E}\left(\text{Var}(X|Y)\right) + \text{Var}\left(\mathbb{E}(X|Y)\right), \tag{5.4}$$

i.e., the sum of the within-variance and the between-variance. Conditioning on the sampling, we get

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{D}\sum_{d=1}^{D} W_d + \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta_D}\right)^2, \tag{5.5}$$

where $\bar{\theta_D} = \frac{1}{D}\sum \hat{\theta}_d$ Finally, equation (5.5) can be rewritten into

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{M}\sum_{i=1}^{M}\widehat{\text{Var}}\left(\hat{\theta}_d^i\right)\right) +$$

$$\frac{1}{D}\sum_{d=1}^{D}\left(\left(1 + \frac{1}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_d^i - \hat{\theta}_d\right)^2\right) +$$

$$\frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta_D}\right)^2.$$

Bibliography

A. Agresti. *Categorical data analysis.* Wiley-Interscience, 2013.

J. M. Alho. Logistic regression in capture-recapture models. *Biometrics*, 46:623 – 635, 1990.

K. Arts, B. F. M. Bakker, and E. Van Lith. Linking administrative registers and household surveys. *Special issue Netherlands Official Statistics*, 15:16 – 22, 2000.

B. Baffour, J. J. Brown, and P. W. F. Smith. An investigation of triple system estimators in censuses. *Statistical journal of the International Association for Official Statistics*, 29:53 – 68, 2013.

B. F. M. Bakker. *Trek alle registers open! (Open all registers!).* Vrije Universiteit Amsterdam, 2009.

B. F. M. Bakker, S. C. Gerritse, P. G. M. van der Heijden, D. J. van der Laan, H. N. van der Vliet, and M. J. L. F. Cruyff. Estimation of non-registered usual residents in the Netherlands, ultimo september 2010. *Conference of European Statistics Stakeholders, Rome, Italy*, 24 - 24 november 2014.

W. R. Bell. Using information from demographic analysis in post-enumeration survey estimation. *Journal of the American Statistical Association*, 88:1106 – 1118, 1993.

Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis.* MIT press, Cambridge, MA., 1975.

I. Boden, L. Capturerecapture estimates of the undercount of workplace injuries and illnesses: Sensitivity analysis. *American Journal Of Industrial Medicine*, 57:1090 1099, 2014.

J. J. Brown, I. D. Diamond, R. L. Chambers, L. J. Buckner, and A. D. Teague. A methodological strategy for a one-number Census in the UK. *Journal of the Royal Statistical Society. Series A*, 162:247 – 267, 1999.

J. J. Brown, O. Abott, and I. D. Diamond. Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 169:883 – 902, 2006.

J. R. Bryant and P. Graham. A bayesian approach to population estimation with administrative data. *Journal of Official Statistics*, 31:475 – 487, 2015.

S. Buuren, van. *Flexible imputation of missing data.* CRC Press, 2012.

Central Bureau of Statistics. StatLine, 15 April 2015.

A. Chao, P. K. Tsay, S.-H. Lin, W.-Y. Shau, and D.-Y. Chao. Tutorial in biostatistics. the application of capture-recapture models of epidemiological data. *Statistics in Medicine*, 20:3123 – 3157, 2001.

L. D. Consiglio and T. Tuoto. Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, 31:415 – 429, 2015.

R. M. Cormack. The statistics of capture-recapture methods. *Oceanography and Marine Biology: An Annual Review 6*, pages 455 – 506, 1968.

R. M. Cormack. *Models for Capture-Recapture Studies, in Statistical Ecology: Sampling Biological Populations.* International Co-operative Publishing House, Fairland, Maryland., 1979.

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

R. M. Cormack, Y.-F. Chang, and G. S. Smith. Estimating deaths from industrial injury by capture-recapture: A cautionary tale. *International Journal of Epidemiology*, 29:1053 – 1059, 2000.

A. A. N. Cruts and M. W. Van Laar. Aantal problematische harddrugsgebruikers in nederland (frequency of problematic harddrug users in the netherlands). *Trimbos Instituut, Utrecht*, 2010.

J. N. Darroch. The multiple-recapture census. ii. estimation when there is immigration or death. *Biometrika*, 46:336 – 351, 1968.

T. De Waal, J. Pannekoek, and S. Scholtus. *Handbook of statistical data editing and imputation.* Wiley, 2011.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1 – 38, 1977.

Y. Ding and S. E. Fienberg. Dual system estimation of Census undercount in the presence of matching error. *Survey Metholodogy*, 20: 149–158, 1994.

European Monitoring Centre for Drugs and Drug Addiction (EMCDDA). Methodological pilot study of local level prevalence estimates. *Lisbon, EMCDDA*, 1997.

European Parliament. Regulation (ec) no 763/2008 of the european parliament and the council of 9 july 2008 on population and housing censuses. *Official Journal of the European Union*, 13.8.2008:L 218/14–L 218/20, 2008.

I. P. Fellegi and A. B. Sunter. A theory of record linkage. *Journal of the American Statistical Association*, 64:1183 – 1210, 1969.

E. Fienberg, S. Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18: 143 – 154, 1992.

S. E. Fienberg. The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, 59:409 – 439, 1972.

S. E. Fienberg and D. Manrique-Vallier. Integrated methodology for multiple system estimation and record linkage using a missing data formulation. *Advanced Statistical Analysis*, 93:49 – 60, 2008.

S. C. Gerritse, B. F. M. Bakker, and P. G. M. van der Heijden. Different methods to complete datasets used for capture-recapture estimation: estimating the number of usual residents in the netherlands. *Statistical Journal of the IAOS*, 31:613627, 2015a.

S. C. Gerritse, P. G. M. Van der Heijden, and B. F. M. Bakker. Sensitivity of population size estimation for violating parametric assumptions in loglinear models. *Journal of Official Statistics*, 31:357 – 379, 2015b.

S. C. Gerritse, B. F. M. Bakker, D. Zult, and P. G. M. van der Heijden. The effects of imperfect linkage and erroneous captures on the population size estimator. *Submitted*, 2016.

M. Gijsberts and M. Lubbers. *Langer in Nederland, ontwikkelingen in de leefsituatie van migranten uit Polen en Bulgarije in de eerste jaren na migratie (Longer in the Netherlands, Developments on living conditions of migrants from Poland and Bulgaria in the first years after migration.).* Sociaal Cultureel Planbureau, The Hague, 2015.

T. N. Herzog, F. J. Scheuren, and W. Winkler. *Data Quality and Record Linkage Techniques.* Springer-Verlag New York, 2007.

E. J. M. Hoogteijling. Raming van het aantal niet in de gba geregistreerden. Technical report, Centraal Bureau voor de Statistiek, 2002.

B. Hook, E. and R. Regal, R. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates. *American Journal of Epidemiology*, 137:1148 – 1166, 1993.

E. B. Hook and R. R. Regal. The value of capture-recapture methods even for apparent exhaustive surveys. *American Journal of Epidemiology*, 135:1060 – 1067, 1992.

E. B. Hook and R. R. Regal. Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews*, 17:243 – 264, 1995.

E. B. Hook and R. R. Regal. Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *American Journal of Epidemiology*, 145:1138 – 1144, 1997.

E. B. Hook and R. R. Regal. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: Internal validity analysis of data from five sources. *American Journal of Epidemiology*, 152:771 – 779, 2000.

V. Hraud-Bousquet, F. Lot, M. Esvan, F. Cazein, C. Laurent, J. Warszawski, and A. Gallay. A three-source capture-recapture estimate of the number of new hiv diagnoses in children in france from 20032006 with multiple imputation of a variable of heterogeneous catchability. *BMC Infectious Diseases*, 12, 2012.

International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology*, 142: 1047 – 1058, 1995.

S. N. Jarvis, P. J. Lowe, A. Avery, S. Levene, and R. M. Cormack. Children are not goldfish - mark/recapture techniques and their application to injury data. *Injury Prevention*, 6:46 – 50, 2000.

J. Kropko, B. Goodrich, A. Gelman, and J. Hill. Multiple imputation for continuous and categorical data: Comparing joint and conditional approaches. *Political Analysis*, 22:497–519, 2014.

G. Lanzieri. Population definitions at the 2010 censuses round in the countries of the UNECE region. *5th meeting of the UNECE Group of*

Experts on Population and Housing Censuses, Geneva, 30 September 3 October 2013.

F. Lawless, J. *Statistics in action. A Canadian outlook.* Apple Academic Press Inc. Ontario, Canada, 2014.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data.* John Wiley and Sons, Hoboken, New Yersey, 2002.

X. L. Meng and D. B. Rubin. Ipf for contingency tables with missing data via the ecm algorithm, in proceedings of the statistical computing section of the american statistical association. *American Statistical Association, Washington, DC.*, page 244247, 1991.

R. Nirel and H. Glickman. Sample surveys and censuses. *Sample surveys: Design Methods and Applications*, 29A:539 – 565, 2009.

ONS. 2011 census item edit and imputation process: 2011 census: Methods and quality report. Technical report, Office for National Statistics, Newport, South Whales, 2012a.

ONS. *The 2011 Census coverage assessment and adjustment process.* Office for National Statistics, Newport, South Whales, 2012b.

D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson. Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62:3 – 135, 1978.

M. Peeters, M. Zondervan-Zwijnenburg, G. Vink, and A. G. J. Van der Schoot. How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology*, 2015.

K. H. Pollock, J. D. Nichols, C. Brownie, and J. E. Hines. Statistical inference for capture-recapture experiments. *Wildlife monographs*, 107: 3 – 97, 1990.

J. L. Schafer. Analysis of incomplete multivariate data. monographs on statistics and applied probability. *Chapman and Hall, London*, 1997a.

J. L. Schafer. Imputation of missing covariates under a general linear mixed model. *Department of Statistics, Pennsylvania State University*, 1997b.

J. L. Schafer. *Analysis of incomplete multivariate data.* Chapman and Hall, 1997c.

G. A. F. Seber. *The Estimation of Animal Abundance and Related Parameters.* Edward Arnold: London, 1982.

C. C. Sekar and W. E. Deming. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44:101–115, 1949.

Statistics Netherlands. *Dutch Census 2011: Analysis and Methodology.* Statistics Netherlands, The Hague/Heerlen, Netherlands, 2015.

J. M. Sutherland, C. J. Schwarz, and L.-P. Rivest. Multilist population estimation with incomplete and partial stratification. *Biometrics*, 63: 910 – 916, 2007.

M. Temurhan, R. Meijer, S. Choenni, M. Van Ooyen-Houben, G. Cruts, and M. van Laar. Capture-recapture method for estimating the number of problem drug users: The case of the Netherlands. *In proceeding of: Intelligence and Security Informatics Conference (EISIC)*, 2011.

United Nations Economic Commission for Europe (UNECE). Practices of UNECE countries in the 2010 round of censuses. *United Nations, New York and Geneva*, 2014.

United Nations Statistics Division. *Principles and Recommendations for Population and Housing Censuses Rev.2.* Statistical papers Series M No 67/Rev.2. United Nations, New York, 2008.

P. G. M. Van der Heijden, M. J. L. F. Cruyff, and G. Van Gils. Aantallen geregistreerde en niet-geregistreerde burgers uit MOE-landen die in Nederland verblijven. Rapportage schattingen 2008 en 2009. (The number of registered and non-registered citizens from MOE-countries residing in the Netherlands. Reporting estimations 2008 and 2009.). *The Hague, Ministry of Social Affairs and Employment*, 2011.

P. G. M. Van der Heijden, J. Whittaker, M. J. L. F. Cruyff, B. F. M. Bakker, and H. N. Van der Vliet. 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.*, 6:831 852, 2012.

K. M. Wolter. Some coverage error models for census data. *Journal of the American Statistical Association*, 81:338 – 346, 1986a.

K. M. Wolter. *Capture-Recapture estimation in the presence of a known sex ratio.* Bureau of the Census Statistical Research Division Report Series, SRD Research Report Number: CENSUS/SRD/RR-86/20, 1986b.

L.-C. Zhang. On modelling register coverage errors. *Journal of Official Statistics*, 31:381396, 2015.

L.-C. Zhang and J. Dunne. *Population size estimation based administrative registers.* 4th Baltic-Nordic Conference on Survey Statistics Conference Proceedings, 24 - 28 August 2015, Helsinki., 2015.

E. N. Zwane and P. G. M. Van der Heijden. Semiparametric models for capturerecapture studies with covariates. *Computational Statistics and Data Analysis*, 47:729 743, 2004.

E. N. Zwane and P. G. M. Van der Heijden. Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registries. *Statistics in Medicine*, 26:1069 – 1089, 2007.

E. N. Zwane and P. G. M. Van der Heijden. Capture-recapture studies with incomplete mixed categorical and continuous covariates. *Journal of Data Science*, 6:557 572, 2008.

6

Summary in Dutch

Susanna C Gerritse

Statistics Netherlands

CONTENTS

6.0.1 Nederlandse samenvatting

Statistische bureaus zoals het Centraal Bureau voor de Statistiek (CBS) worden periodiek gevraagd om aan te geven hoeveel inwoners een land bevat. Het aantal inwoners van een land wordt in het Engels aangeduid met usual residents, ofwel de gewoonlijke bevolking. Volgens EU regelgeving wordt de gewoonlijke bevolking gedefinieerd als de mensen die langer dan 12 maanden in een land verblijven, of de intentie hebben dat te doen. In Nederland maakt het Centraal Bureau voor de Statistiek gebruik van een Populatie Register (PR) om het aantal inwonenden te schatten: de Gemeentelijke basisadministratie (GBA) en vanaf 1 januari 2014 de Basisregistratie personen (BRP). De Nederlandse wet schrijft voor dat mensen zich moeten inschrijven in de PR als zij langer dan 4 maanden verblijven in Nederland, of de intentie daartoe hebben. Echter, de PR alleen is niet afdoende om de omvang van de gewoonlijke bevolking te schatten, aangezien er ook mensen zijn die wel in Nederland wonen, of de intentie hebben voor minstens 12 maanden in Nederland te wonen, maar die zich niet hebben ingeschreven. Twee zulke grote groepen zijn voormalige asielzoekers die uitgeprocedeerd zijn, en Europese arbeidsmigranten die wegens vrij verkeer van personen niet illegaal in Nederland verblijven als ze zich niet inschrijven in het GBA. Om de gewoonlijke bevolking te schatten heeft de PR daarom een onderdekking, en een oplossing moet gevonden worden om deze onderdekking te schatten.

Vangst-hervangst methodologie word gebruikt om te schatten welk deel van de gewoonlijke bevolking wordt gemist door de PR. Als twee of meer registers aan elkaar gekoppeld worden, kan met de vangst-hervangst methode worden geschat hoeveel mensen in deze registers zijn gemist terwijl zij wel tot de populatie behoren. In dit proefschrift worden een politieregister, de HerKenningsdienst Systeem (HKS), en de werknemersdeel uit de Polisadministratie, het WerkNemersBestand (WNB), gekoppeld aan de PR om te schatten welk deel van de populatie deze registers tezamen missen om zodoende tot een onderdekking te komen van de PR.

De vangst-hervangst methodologie is ontwikkeld om dierpopulaties te schatten en is later uitgebreid om aantallen mensen te schatten. In het simpelste voorbeeld worden twee registers gebruikt, waarbij mensen in ieder van de registers wel of niet zijn opgenomen. Hierdoor kunnen aantallen mensen worden geclassificeerd in een 2 bij 2 tabel. Echter, het aantal mensen dat in beide registers niet voorkomt maar wel tot de bevolking hoort, is een zgn. structurele nul in de 2 bij 2 tabel, aangezien deze mensen per definitie niet zijn geobserveerd in beide registers. Deze structurele nul wordt bijgeschat door toepassing van de methode. Er zijn uitbreidingen van de methode naar meer dan twee registers en naar covariaten. De methode werkt goed als er aan een aantal aannames is gedaan. Echter, deze aannames worden vaak geschonden. Vanuit de geobserveerde data kan niet worden geverifieerd of de aannames geschonden zijn, laat staan in welke mate.

De registers die worden gebruikt zijn aan elkaar gekoppeld door gebruik te maken van een combinatie van deterministisch en probabilistisch koppelen. Op het CBS worden alle registers standaard deterministisch gekoppeld aan de PR. Om deterministisch te koppelen moeten twee individuen op een zgn. koppelsleutel exact overeenkomen, waarbij enkel kleine fouten zoals spelfouten nog meegenomen kunnen worden. Echter, het zijn vooral de mensen die in Nederland verblijven zonder Nederlandse nationaliteit die zich vaak ook niet hebben geregistreerd als inwoner. De van deze mensen in de registers aanwezige informatie bevat ook vaak registratiefouten en daarom is het lastig om hen deterministisch te koppelen aan de PR. Zodoende maken we ook gebruik van probabilistisch koppelen, om de kans te vergroten om alle individuen in het HKS en de WNB te vinden die in de PR zijn ingeschreven.

Om tot een schatting te komen van de omvang van de gewoonlijke bevolking, moeten aantallen personen worden uitgesplitst naar verblijfsduur. Dat kan door verblijfsduur als covariaat in het model dat gebruikt wordt voor de vangst-hervangst schatting op te nemen. Echter, de drie gebruikte registers hebben geen kant-en-klare variabele voor verblijfsduur. In de PR is een dag van inschrijving bekend, waaruit verblijfsduur kan worden afgeleid. In de WNB zijn baanduren bekend en deze kunnen worden gebruikt om tot een schatting te komen van verblijfsduren. In de HKS moet een schatting voor verblijfsduur worden ingevuld, geimputeerd, gebruikmakend van de informatie van de andere twee bronnen.

In dit proefschrift worden er twee hoofdvragen beantwoord: 1) Wat is het effect van schending van aannames en ontbrekende waarden op de populatieschatter, en 2) hoe kun deze informatie worden gebruikt om tot een accurate, realistische schatting te komen voor de onderdekking van de PR? Hoofdstuk 1 geeft achtergrondinformatie over het proefschrift. Hoofdstukken 2 tot en met 4 gaat in op hoofdvraag nummer 1, en hoofdstuk 5 gaat in op hoofdvraag 2. En term overkoepelt het hele proefschrift en dat is implied coverage, de dekking van register 1 die wordt gempliceerd door andere registers. Gempliceerde dekking wordt kort besproken in hoofdstuk 2, maar pas in hoofdstuk 3 is beschreven dat de gempliceerde dekking van registers een grote invloed heeft op de mate waarin geschonden aannames van effect zijn op de populatieschatter.

In hoofdstuk 2 wordt de gevoeligheid besproken van populatieschatters wanneer bepaalde parametrische aannames worden geschonden. In het bijzonder wordt in dit hoofdstuk schending van de aanname van onafhankelijkheid besproken. Onder onafhankelijkheid nemen we aan dat de insluitkansen van register 1 onafhankelijk zijn van de insluitkansen van register 2. Omdat we onafhankelijkheid niet kunnen verifieren uit de data, is er een methode bedacht om afhankelijkheid te simuleren en het effect hiervan op de schatter te onderzoeken. In dit hoofdstuk worden er twee nationaliteitsgroepen vergeleken: de eerste groep bevat mensen met een Poolse nationaliteit en de tweede groep bevat mensen met een Afghaanse, Iraakse of Iraanse nationaliteit. Bij simulering van afhankelijkheid bleek voor de individuen met een Poolse nationaliteit dat de uitkomsten een grote spreiding te zien gaven. Daarom kan, bij onzekerheid over de mate van schending van de onafhanklijkheidsaanname, een schatting niet zomaar vertrouwd worden. De schatting voor de mensen met een Afghaanse, Iraaks en Iraanse nationaliteit was wel robuust onder gesimuleerde afhankelijkheid. Deze conclusie bleef overeind bij toevoeging van een covariaat (geslacht). Ook wordt er in dit hoofdstuk besproken hoe men ten dele geobserveerde covariaten toch kan gebruiken als covariaat in de vangst-hervangst schatting. Tussen de deels geobserveerde covariaat en het register werd afhankelijkheid gesimuleerd, waarbij voor beide nationaliteiten de schatter robuust bleef.

In hoofdstuk 3 wordt de gevoeligheid besproken van de populatieschatter onder simulering van foutieve koppelingen en foutieve vangsten. In de vangst-hervangst methode wordt aangenomen dat er perfect kan worden gekoppeld en dat er geen foutieve vangsten zijn, zodat elk individu in de registers tot de populatie behoort. Dezelfde twee nationaliteitsgroepen werden gebruikt, maar dan voor zowel het twee register model als het drie register model. Weer werd gevonden dat de schatter voor de Poolse groep niet robuust was onder geschonden aannames, en deze voor de groep Afghanen, Irakezen en Iranezen wel. Er werd onderzocht waarom dit zo is. Voortbouwend op eerder uitgevoerd onderzoek, is de term geimpliceerde dekking in dit hoofdstuk verder uitgewerkt. In het geval van twee registers, is de geimpliceerde dekking hoog als register 2 weinig nieuwe personen toevoegt aan register 1. Als deze dekking hoog is, dan is de schatter robuust voor schendingen van de aannames. Is de dekking echter laag, dan is de schatter niet robuust en is de schatting onbetrouwbaar.

In hoofdstuk 4 wordt besproken hoe er omgegaan kan worden met ontbrekende gegevens (missing values) op de verblijfsduur variabele. Verblijfsduur is van cruciale betekenis voor dit onderzoek, daarom moest er een oplossing worden gevonden voor de ontbrekende gegevens. Voor de ontbrekende gegevens zijn op vier manieren, scenarios, waarden ingevuld en de resulterende populatieschattingen zijn met elkaar vergeleken. Drie van de vier scenarios gebruikten het Expectation Maximization (EM)-algoritme, en het laatste scenario maakte gebruik van Predictive Mean Matching (PMM). De uitkomsten van de vier scenarios zijn met elkaar vergeleken om te besluiten welke methode het beste de missende waardes kan voorspellen en de beste uitkomst geeft. Om tot dit besluit te komen is er allereerst literatuuronderzoek gedaan naar eerder onderzoek met vergelijkbare schattingen, waaruit een verwachting kon worden afgeleid voor de schatting voor de cijfers uit 2010. Daarnaast gebruikten we informatie over de asielaanvragen in 2010 en de jaren ervoor. De verwachte schatting voor 2010 zou moeten liggen tussen 175 en 225 duizend personen, en alleen de schatting voor de PMM-imputatie viel daarbinnen. Methodologisch was er ook een voorkeur voor de PMM-imputatie omdat dit de enige methode is die de mogelijkheid biedt slechts een deel van de geobserveerde data te gebruiken voor het imputeren in plaats van alle geobserveerde data. De schatting kwam er op uit dat 212 duizend individuen worden gemist door de PR . Let wel, deze schatting had alleen betrekking op mensen met een leeftijd tussen de 15 en 65, vanwege de begrenzing van de registers. De schatting moet nog worden aangevuld met een schatting van mensen jonger dan 15 en ouder dan 65.

Hoofdstuk 5 gebruikt de informatie die is verkregen uit de eerdere hoofdstukken en ander voorgaand onderzoek om tot een accurate schatting te komen van de onderdekking van de GBA om het aantal personen te schatten dat deel uit maakt van de gewoonlijke bevolking. Dit hoofdstuk kan als voorbeeld worden gebruikt van hoe er omgegaan kan worden met de praktische problemen die men aantreft bij het schatten van de bevolkingsgrootte. Zo wordt elke aanname afzonderlijk besproken, alsmede wat er is gedaan om aan deze aannames te voldoen. Het bleek dat 37% van de individuen in de HKS die niet koppelen aan de andere registers, incomplete koppelsleutels hadden, waardoor deze mensen niet gekoppeld konden worden en niet bepaald kon worden of het hier wellicht om foutieve vangsten ging. Een aantal scenarios wordt gebruikt om te beoordelen hoe groot de spreiding is van de geschatte uitkomsten. Uiteindelijk komen we tot een schatting die ligt tussen 88 tot 186 duizend individuen die gemist worden door de PR, wat neerkomt op een onderdekking van maar 0.5 tot 1.1%. Ook hier moet worden opgemerkt dat het gaat om de mensen tussen 15 en 65 jaar, en de onderdekking van de PR zal iets hoger liggen als ook kinderen tot 15 jaar en ouderen vanaf 65 jaar worden meegenomen.

In dit proefschrift is besproken hoe de vangst hervangst methode gebruikt kan worden om tot een schatting te komen van de gewoonlijke bevolking van Nederland. Er is onderzocht hoe de aannames en de gempliceerde dekking van invloed zijn op de schatter. Eveneens is er onderzocht hoe er omgegaan kan worden met mogelijke praktische problemen zoals missende waardes. Het proefschrift draagt hiermee bij aan algemene kennis over de vangst-hervangst methode alsmede over schatten binnen de officile statistiek.

Curriculum Vitae

Susanna Charlotte Gerritse was born to Henk and Louise Gerritse in Amsterdam in 1988, raised in Almere where her brother Frank was born. After completing a VWO studies in Almere she returned to Amsterdam to study psychology. She finished her Bachelor in Clinical Psychology and a Research Master Psychology at the University of Amsterdam. During her master she switched to a major in methodology instead of Clinical Psychology. She interned for Professor Dr. Denny Borsboom to investigate a network perspective for schizophrenia symptoms and did her masters thesis on Catastrophy modeling of addiction data for Professor Dr. Han van der Maas.

Intrigued by methodology and statistics, she started a phd, of which the resulting Thesis is now in your hands. She has presented her work on a national level but also international, including being invited to present at the Central Statistics Office in Ireland, Istat in Italy, and conferences in Belgium and Hong Kong. She was selected to attend the Doctoral Summerschool on Ethics and Integrity by the League of European Research Universities (LERU) in Helsinki, Finland.

During her PhD Susanna was active as a board member for Prout (Promovendi Overleg Utrecht) and PNN (Promovendi Netwerk Nederland). Here she was active to represent the PhDs at Utrecht University and later on a national level. Amongst others, she gave the Professional PhD Program of the PNN an extra boost and contributed to the discussions on bursary PhDs.

Currently she is a statistical researcher at Statistics Netherlands. She moved with her man to The Hague where they live with her two cats.

"The thing I treasure most in life, cannot be taken away
There will never be a reason why, I will surrender to your advice
To change myself, I'd rather die
Though they will not understand
I won't make the greatest sacrifice
You can't predict where the outcome lies
You'll never take me alive
I'm alive!"

Disturbed - I'm alive