SAPIENZA
UNIVERSITÀ DI ROMA

Data Integration and Official Statistics, with a focus on Bayesian models for population size estimation

Doctoral School of Economics

Dottorato di Ricerca in Models for Economics and Finance – XXXIV Ciclo

Candidate

Tiziana Tuoto
ID number 950561

Thesis Advisor

Prof. Andrea Tancredi

Co-Advisor

Dr. Davide Di Cecco

2021/2022

Data Integration and Official Statistics, with a focus on Bayesian models for population size estimation
Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: tuoto@istat.it

Abstract

The integrated use and the re-use of data coming from different sources is a common practice in official statistics and it is recognized by the international community as a key element of modernization of the statistical system. Actually, data generated for purposes other than statistical can often be easily acquired at a low cost, hence data integration reduces the costs of data collection and limits the statistical burden on the respondents. In this research project, we have developed three different aspects related to data integration activities in official statistics.

In Chapter 1 we considered the use of data from administrative archives to support survey data on a sensitive variable, income. This research was carried out in cooperation with Prof. Li-Chun Zhang, from Statistics Norway, University of Southampton, and Olso University, during his frequent visits to Rome, at the National Statistical Institute (Istat) and Sapienza University. We assumed that a data linkage has been performed to combine administrative data and survey data with the aim of identifying and bringing together records from separate files, which correspond to the same entities. Usually, data linkage is not a trivial procedure and linkage errors, false and missed links, might affect standard statistical techniques, producing misleading inference. In this setting, we developed a regression model on integrated data for secondary analysis, where the linked data has been prepared by someone else, and neither the match-key variables nor the unlinked records are available to the analyst. We developed also a diagnostic test for the assumption of non-informative linkage errors, which is required for our proposal as well as for all existing secondary analysis adjustment methods. Compared to other adjustment methods, our approach provides important advantages: it relies on a realistic assumption that the probabilities of correct linkage vary across the records but it does not assume that one is able to estimate the probability of correct linkage for each individual record. Moreover, it accommodates in a simple manner the general situation where the files are of different sizes and none of them is a subset of another. The adjusted regression model and the proposed test have been studied by simulation and also applied to real data. The research illustrated in Chapter 1 has published as original article by the Journal of the Royal Statistical Society: Series A (Statistics in Society) Volume 184, Issue 2.

In Chapter 2, we dealt with a different data integration problem. We considered an additional re–use of an administrative register on prosecuted crimes to estimate the size of certain criminal populations, and in particular the size of those involved in criminal activities but for some reasons unreported to the justice system. In the capture-recapture framework of repeated count data, we focused on the identification and treatment of "one–inflation". This phenomenon occurs when the number of units captured exactly once clearly exceeds the expectation under a baseline count distribution. It has received increasing attention in capture–recapture literature in recent years, since ignoring one–inflation has serious consequences for the estimation of the population size, which can be drastically overestimated. Criminal data might be particularly prone to the one–inflation, since people involved might develop an extreme form of the so–called "trap shy" behavioural model, i.e. the will and ability to avoid subsequent captures. In Chapter 2, in a joint work with supervisors Prof. Andrea Tancredi and Dr. Davide Di Cecco, we proposed a Bayesian approach for Poisson, Geometric and Negative Binomial one–inflated count distributions. Posterior inference for population size is obtained applying a Gibbs sampler approach. We also provided a Bayesian approach to model selection. We illustrated the proposed methodology with simulated and real data to estimate the number of people implicated in the exploitation of prostitution in Italy. The research illustrated in Chapter 2 has been published as research article by Biometrical Journal Volume 64, Issue 5, in March 2022.

In Chapter 3 we extended the models presented in Chapter 2 in two directions. From one side, we distinguished at least two possible causes for one–inflation, namely, the erroneous inclusion of out–of–scope units, and the behavioral effect preventing subsequent captures after the first one. Accordingly, we propose two families of one–inflated models to estimate the number of uncaptured units. In addition, we proposed a Bayesian semi-parametric approach by considering a Dirichlet process mixture model as a base model, and extend this class to include one–inflation, in order to take into account also the unobserved heterogeneity in the captures probabilities. We also compare the Dirichlet process mixture models with sparse finite mixture (SFM) models which, to the best of our knowledge, even if strictly related to DPMs, have not yet been applied in capture–recapture field. The mixture models and the two one–inflated counterparts were compared on three datasets of criminal proceedings. Chapter 3 is the result of a yet unpublished work by myself, and my supervisors Prof. Andrea Tancredi and Dr. Davide Di Cecco.

Contents

Chapter 1

Linkage-data linear regression

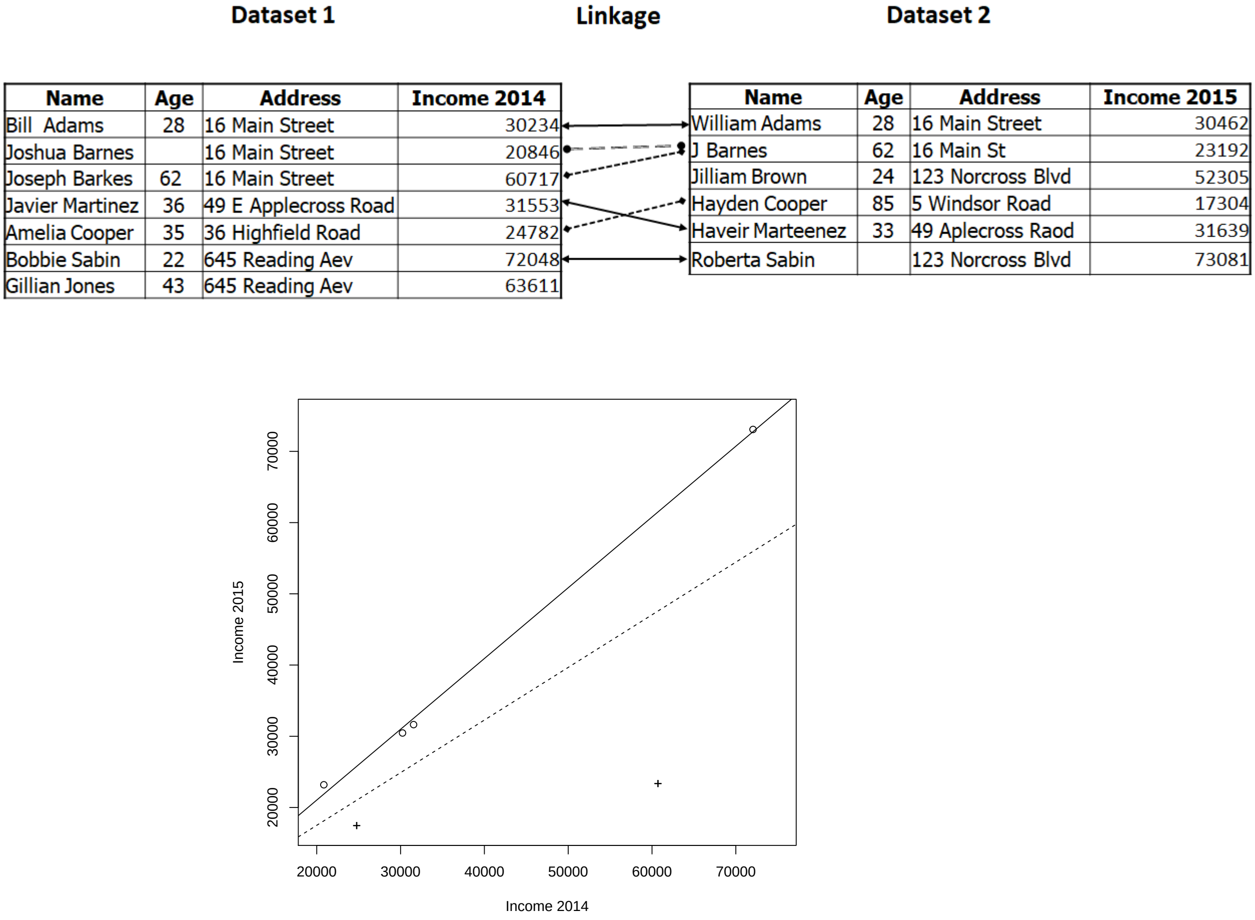
Abstract

Data linkage is increasingly being used to combine data from different sources with the aim of identifying and bringing together records from separate files, which correspond to the same entities. Usually, data linkage is not a trivial procedure and linkage errors, false and missed links, are unavoidable. In these cases, standard statistical techniques may produce misleading inference. In this Chapter, we propose a method for secondary linear regression analysis, where the linked data has to be prepared by someone else, and neither the match-key variables nor the unlinked records are available to the analyst. We develop also a diagnostic test for the assumption of non-informative linkage errors, which is required for all existing secondary analysis adjustment methods. Our approach provides important advantages: it relies on the realistic assumption that the probabilities of correct linkage vary across the records but it does not assume that one is able to estimate the probability of correct linkage for each individual record. Moreover, it accommodates in a simple manner the general situation where the files are of different sizes and none of them is a subset of another. The proposed methodology of adjustment and testing is studied by simulation and applied to real data.

1.1 Introduction

Computerised record linkage is increasingly common for scientific investigation, policy analysis and commercial development, where one aims to identify and bring together the records (with associated observations) in separate data files, which correspond to the same entities or individuals (Fellegi & Sunter (1969); Herzog et al. (2007); Christen (2012); Harron et al. (2015)). Industrial-strength applications to large population-size datasets have become relatively straightforward, e.g. when population census data files are linked over time to create longitudinal population datasets (Zhang & Campbell (2012)), or population-wide administrative registers are linked to create pseudo population spine in the absence of a Central Population Register (Owen et al. (2015)). In epidemiology and medical studies, record linkage is extensively used in many countries to enhance data on clinical performance and patient health outcomes (e.g. Harron et al. (2016)). Record linkage is a necessary step for estimating the size of hidden or hard-to-count populations, i.e. illegal drug users, drinking drivers, illegal migrants, civil war victims, just to cite few examples of studies on human population (Rosman (2001); van der Heijden et al. (2014); Seybolt et al. (2013)); studies on wild animal populations provide plenty of application (Creel et al. (2003); Link et al. (2010); McClintock et al. (2014); Wright et al. (2009)). In our illustrative application in Section 1.4, we consider linked income data from tax registers in two consecutive years, and linear regression of year-on-year incomes for a simple analysis of the development at local (municipality) level. Using administrative data here allows for disaggregated analysis that otherwise cannot be supported by survey sampling, because of the limited sample size and the fact that income may be considered "sensitive", which causes non-response and/or under-reporting errors in surveys.

When there does not exist a unique identifier that allows for exact matching, record linkage is performed using soft identifiers, the so-called *key variables*, such as name, age, address, etc. Let each pairing of records of the same entity be a *match*. Let each pairing of records that results from record linkage be a *link*. Insofar as the key variables may be affected by measurement errors, *linkage errors* are unavoidable, so that the links may not be identical to the matches. There are two types of linkage error: either the linked records do not actually refer to the same entity, or if one fails to link the records that refer to the same entity. Figure 1.1 provides an illustration using fictive individuals and income data, where there are three correct *links* (solid), two errors of *false linkage* (dashed) and one of



Dataset 1

Name	Age	Address	Income 2014
Bill Adams	28	16 Main Street	30234
Joshua Barnes		16 Main Street	20846
Joseph Barkes	62	16 Main Street	60717
Javier Martinez	36	49 E Applecross Road	31553
Amelia Cooper	35	36 Highfield Road	24782
Bobbie Sabin	22	645 Reading Aev	72048
Gillian Jones	43	645 Reading Aev	63611

Linkage

Dataset 2

Name	Age	Address	Income 2015
William Adams	28	16 Main Street	30462
J Barnes	62	16 Main St	23192
Jilliam Brown	24	123 Norcross Blvd	52305
Hayden Cooper	85	5 Windsor Road	17304
Haveir Marteenez	33	49 Aplecross Raod	31639
Roberta Sabin		123 Norcross Blvd	73081

Figure 1.1. Fictive income data 2014-2015. On the top: record linkage, the lines represent correct links (), false links (), and missing match (); on the bottom: linear regression, () based on unknown matches (o), and () based on observed links (+).

missing match (long-dashed). The plot shows the ordinary least squares fit (solid line) based on the four unknown matches (circle) and that (dashed) based on the five observed links ("+" for the two incorrect links). Clearly, treating the linked dataset as if it were true generally causes bias of the resulting analysis. For a situation like the one in Figure 1.1, one needs to deal with *at least* three problems.

- Different individuals (or entities) can have different probabilities for being incorrectly linked or missed (given a match exists), which we refer to as the problem of *heterogeneous linkage errors*.

- There are *unmatched* individuals in *both* files that cannot possibly be correctly linked, which we refer to as the problem of *incomplete match space*. In Figure 1.1 these are Barkes, A. Cooper and Jones in file 1, and Brown and H. Cooper in file 2. *Complete match space* would have been the case here had none of these unmatched individuals existed, or if they had only existed in *one* of the two files, say, when file 1 is a sample taken from file 2.

- Whether (Joshua Barnes, J. Barnes) are a match is a mutually exclusive event of whether (Joseph Barkes, J. Barnes) are a match, as long as there are no duplicated records in each file, which we refer to as the problem of bi-partite *linkage data structure*. Due to the bi-partite linkage data structure (Jaro (1989)), it would e.g. be wrong to model (Joshua Barnes, J. Barnes)'s being a match as a Bernoulli event that is statistically independent of (Joseph Barkes, J. Barnes)'s being a match.

1.1.1 Related works

The awareness of misleading inference from standard statistical techniques in the presence of linkage errors dates back to Neter et al. (1965). Linear regression is studied by Scheuren & Winkler (1993), Scheuren & Winkler (1997), and Lahiri & Larsen (2005), where the data analyst and the linker are essentially the same. Chambers (2009) and Chambers & da Silva (2019) adopt the perspective of secondary analysts, who have no access to the key variables and the separate data files, nor the detailed knowledge or tools to replicate the actual linkage procedure (Zhang (2019)). Consequently, Chambers (2009) adopts a greatly simplifying assumption, referred to as the *exchangeable linkage error (ELE)* model, where there exists a constant false linkage probability and mismatching is completely random in the case of false linkage. While the ELE assumption is practically appealing, it cannot properly accommodate heterogeneous linkage errors. Moreover, as we shall explain in more details in Section 1.2, the ELE model is only applicable if one treats any incomplete match space as if it were complete. But the false linkage error of an unmatched individual (such as Brown in Figure 1.1) always has probability one, so it cannot be the same as that of a matched individual (such as Martinez) who can be linked correctly.

Nearly all the frequentist methods for the analysis of linked data are based on the *linkage model* of the probability that a record in one dataset is linked to *each* of the records in the other. The ELE model is the simplest linkage model. Techniques such as regression analysis, estimation equation and analysis of contingency tables are studied by Scheuren & Winkler (1993), Scheuren & Winkler (1997), Lahiri & Larsen (2005), Chambers (2009), Chipperfield et al. (2011), Hof & Zwinderman (2012), Kim & Chambers (2012), Chipperfield & Chambers (2015), Han & Lahiri (2018) and Enamorado et al. (2019). Again, as we shall explain in Section 1.2, in reality the linkage model cannot cope with incomplete match space, even when the ELE assumption is relaxed to accommodate heterogeneous linkage errors. Yet incomplete match space is generally the case when data originate from different sources, such as when linking hospital patient records to welfare payment records. It is fundamentally different to the situation, where one set of individuals form a sample of the other set (i.e., population), where there are no individuals in the sample who cannot possibly be linked correctly.

Bayesian inference is based on the posterior distribution of the unknown set of matched entities. Different modelling approaches are used for the linkage key variables that are subjected to measurement errors, e.g. Tancredi & Liseo (2011) and Steorts et al. (2017) extend the hit-miss model of Copas & Hilton (1990), whereas Sadinle (2014), Sadinle (2017) models the comparison vector of key variables following the Fellegi & Sunter (1969) tradition. See also Gutman et al. (2013) for a modelling approach, which includes both variables subjected to measurement errors and others that do not. However, it is common that the variables being modelled for linkage are inaccessible to the secondary analyst. Handing out multiple posterior sets of matched entities may be impractical, together with the associated variables needed for analysis, especially if the analysis requires a large number of posterior draws. Although there are improvements in the direction of scalability (Marchant et al. (2019)), there still does not exist any reported Bayesian linkage application to files of the size of a population census.

Goldstein et al. (2012) and Gutman et al. (2015) apply multiple imputation techniques to analysis of linkage data, which do not handle the problem of linkage data structure like the other Bayesian methods above. Restrictions due to linkage data structure are not built into these imputation methods.

1.1.2 Outline of the chapter

In this Chapter we consider *linkage-data linear regression*, where one aims to estimate the regression coefficients *only* based on the linked dataset. In particular, we adopt the *secondary* analyst perspective, where the linked data have to be prepared by someone else; neither the unlinked records nor the key variables in the separate files are available to the analyst. We develop a novel frequentist method of *Pseudo Ordinary Least Squares (OLS)*, which deals with all the three problems exemplified above in Figure 1.1, i.e., heterogeneous linkage errors, incomplete match space and linkage data structure. Like all the methods referenced in this Introduction, the key assumption to our approach is that the linkage errors are non-informative of the regression model parameters. The assumption will be defined and discussed in Section 1.2. Moreover, for the first time we shall construct an accompanying diagnostic test for the non-informative linkage-error assumption, which can provide helpful guidance in practice. Application to real income data and simulation studies suggest that the assumption can be met at least approximately in many situations, and the Pseudo-OLS estimator is more efficient than the existing methods in the cases of incomplete match space that are examined here.

The rest of the Chapter is organised as follows. In Section 1.2 we start by introducing the basic notations and the set-up of linkage-data linear regression. In Section 1.2.1, we recall the existing frequentist methods and explain carefully why they do not fully meet the challenges of incomplete match space. Section 1.2.2 defines and discusses the non-informative linkage-error assumption. Our proposed approach is then developed in Sections 1.2.3, 1.2.4 and 1.2.5, including the underlying assumptions and the consistency of the resulting regression coefficient estimator. Section 1.2.6 analyses the bias of the existing methods, which arises from treating the incomplete match space as if it were complete. In Section 1.3 we develop a diagnostic test for the non-informative linkage error assumption. An application to linked income data from tax registers is given in Section 1.4, which demonstrates considerable efficiency gains by our method against the existing ones. We carry out a simulation study in Section 1.5, which helps us to better appreciate the application results and to explore some other aspects of the proposed methodology of adjustment and testing. We conclude with some brief remarks in Section 1.6.

1.2 Methods

Let $y_i = x_i^\top \beta + \epsilon_i$ be a linear regression model, where x_i is the $p \times 1$ vector of covariates, and β is the parameter of interest. Let dataset D_1 contain the covariates x_i for record $i \in D_1$, and let dataset D_2 contain the dependent variable y_j for $j \in D_2$. We assume that duplicated records have been successfully removed from both. Let $N_1 = |D_1|$ and $N_2 = |D_2|$ be the sizes of D_1 and D_2. Let D_M be the set of matched entities between D_1 and D_2, i.e. those ones that can possibly be correctly linked, to which the linear regression model applies. Let

$$\Omega = D_1 \times D_2 = M \cup U,$$

where $M = \{(i,i) : i \in D_M\}$ contains the matches, and U contains all the mismatched pairs of records. Let $N_M = |M|$ be the size of M. In the ideal case, one would estimate β based on the pairs of records in M. However, M is unknown. Suppose a record linkage procedure yields the set of links, between records in D_1^* from D_1 and D_2^* from D_2, respectively, denoted by

$$M^* = \{(i,j) : i \in D_1^*, j \in D_2^*\},$$

where $N^* = |D_1^*| = |D_2^*| = |M^*| \leq \min(N_1, N_2)$, and $M^* \neq M$ whenever linkage errors are present. In linkage-data linear regression one aims to estimate β only based on the linked dataset, which can take on any of the following expressions in this Chapter:

$$(x, y)_{M^*} = \{(x_i, y_j) : (i, j) \in M^*\} = \{(x_i, y_i^*) : y_i^* = y_j,\, (i, j) \in M^*\}.$$

Let $D_{1M}^* = D_1^* \cap D_M$ be the set of matched entities in D_1 that are linked, and $D_{2M}^* = D_2^* \cap D_M$ those from D_2. Let D_{MM}^* be the set of correctly linked entities, where $\{(i, i) : i \in D_{MM}^*\} = M^* \cap M$. Let $N_{MM}^* = |D_{MM}^*|$ be its size. We have $D_{MM}^* \subseteq D_{1M}^* \subseteq D_1^*$ and $D_{MM}^* \subseteq D_{2M}^* \subseteq D_2^*$.

For an illustration using Figure 1.1, let file 1 contain $D_1 = \{1, 2, 3, 4, 5, 6, 7\}$ and let file 2 contain $D_2 = \{1, 2, 8, 9, 4, 6\}$, both in the running order from top to bottom, where $D_M = \{1, 2, 4, 6\}$ are the matched individuals and $\{3, 5, 7, 8, 9\}$ are the unmatched ones. We have $D_1^* = \{1, 3, 4, 5, 6\}$ and $D_{1M}^* = \{1, 4, 6\}$, $D_2^* = \{1, 2, 9, 4, 6\}$ and $D_{2M}^* = \{1, 2, 4, 6\}$. The set of links is $M^* = \{(1, 1), (3, 2), (4, 4), (5, 9), (6, 6)\}$. The correctly linked individuals can only come from D_M, which are $D_{MM}^* = \{1, 4, 6\}$.

1.2.1 Two linkage-model estimators for complete match space

Consider the case of complete match space, where $N \equiv N_1 = N_2 = N_M$. Suppose each record in D_1 is linked to one and only one record in D_2, such that $(N^*, D_1^*, D_2^*) = (N, D_1, D_2)$. The linked y-value for $i \in D_1^*$ is $y_i^* = \sum_{j \in D_2^*} a_{ij} y_j$, where $a_{ij} = 1$ if $i \in D_1^*$ is linked to $j \in D_2^*$ and $a_{ij} = 0$ otherwise. Notice that i and j refer to distinctive records themselves, regardless how they appear or are arranged in the two files. False linkage of $i \in D_1^*$ is the case if $a_{ij} = 1$ for $j \in D_2^*$ and $j \neq i$. However, a_{ij} is unobserved, since the true matches are unknown. What is observed is whether or not $i \in D_1^*$, i.e. record $i \in D_1$ is linked or not, denoted by $\ell_i = 1$ or $\ell_i = 0$. In the special setting here, we have $\ell_i = 1$ for all $i \in D_1$. Denote the conditional expectation of a_{ij} given linkage by

$$p_{ij} = E(a_{ij} | \ell_i = 1) = \Pr(a_{ij} = 1 | \ell_i = 1).$$

Let $P_{N \times N} = [p_{ij}]$ be the matrix of p_{ij}'s. Let $X_{N \times p}$ be the covariate matrix associated with D_1, and $y_{N \times 1}$ the dependent vector of D_2, in the matched ordering such that the diagonal of P corresponds to M. Let $y_{N \times 1}^*$ be the vector of linked y-values, which is a linear transformation of y via $[a_{ij}]$.

Provided the linkage indicators $[a_{ij}]$ are independent of $(x, y)_M$, we have

$$E(y^* | X, y) = Py.$$

Given complete match space, the regression model applies to all the units in D_2, so that $E(y|X) = X\beta$. Thus, $E(y^*|X) = Z\beta$ for $Z = PX$. Lahiri & Larsen (2005) propose OLS fit:

$$\widehat{\beta}_{LL} = (Z^\top Z)^{-1} Z^\top y^*.$$

Chambers (2009) notices in addition an unbiased *adjusted least squares* fit:

$$\widehat{\beta}_A = (X^\top P X)^{-1} X^\top y^*$$

The matrix P does not contain sensitive information and, in theory, could be supplied by the data linker. In practice, however, there is currently a lack of consensus on how to estimate the matrix P. See discussions of alternative approaches in Lahiri & Larsen (2005), Han & Lahiri (2018), Chambers & Kim (2015), and Tuoto (2016). Moreover, these methods require access to the key variables, which is only possible for the data linker. Chambers (2009) proposes the ELE model of P, where

$$p_{ii} = \lambda \qquad \text{and} \qquad p_{ij} = (1-\lambda)/(N-1). \tag{1.1}$$

which ignores the problems of heterogeneous linkage errors. Even when the model (1.1) is relaxed to accommodate heterogenous linkage errors with varying p_{ij}'s, the linkage-model approach still cannot cope with the problem of incomplete match space. In this Chapter, we propose to estimate linear regression parameters when the matrix P is not provided.

Again, take the example in Figure 1.1 and consider Adams ($i = 1$) and Barkes ($i = 3$). The expectation of their linked y-value, respectively, are given as

$$E(y_1^*|X, y, \ell_1 = 1) = p_{11}y_1 + p_{12}y_2 + p_{18}y_8 + p_{19}y_9 + p_{14}y_4 + p_{16}y_6,$$
$$E(y_3^*|X, y, \ell_3 = 1) = p_{31}y_1 + p_{32}y_2 + p_{38}y_8 + p_{39}y_9 + p_{34}y_4 + p_{36}y_6,$$

provided non-informative linkage errors. Since Adams is a matched individual that can be linked correctly, one can e.g. let $p_{11} = \lambda_1$ and $p_{1j} = (1-\lambda_1)/4$, for any other $j \in D_2^*$, given that the secondary analyst only sees the five links that are provided. But this would mean to assume that the unlinked individual Brown ($i = 8$) in $D_2 \setminus D_2^*$ has no chance of being linked with Adams, i.e. treating the incomplete match space as if it were complete. Next, since Barkes is an unmatched individual, it would be totally wrong to act similarly, because there is no record at all in D_2 for Barkes. One might consider setting $p_{3j} \equiv 1/5$ as an assumption of random false linkage. However, without knowing the true matched or unmatched status of Adams and Barkes, one would not know if p_{1j}'s or p_{3j}'s should be assigned. This shows that the linkage-model approach cannot cope with incomplete match space.

Thus, in reality, one can only apply the ELE model (1.1), by assuming that the linked sets (D_1^*, D_2^*) form complete match space in any case. Clearly, this is not satisfactory conceptually: although one may assume $y_3 = x_3^\top \beta + \epsilon_3$ for Barkes in D_1^*, one would not find y_3 among $y^* = (y_1, y_2, y_9, y_4, y_6)^\top$. Similarly, although one may assume that there exists x_9 for Cooper in D_2^*, such that $y_9 = x_9^\top \beta + \epsilon_9$, one would not find x_9 in $X_{D_1^*}$. However, as we will discuss later in Section 1.2.6, doing so may still yield useful bias reduction compared to the *face-value* OLS, given by

$$\widehat{\beta}^* = (X_{D_1^*}^\top X_{D_1^*})^{-1} X_{D_1^*}^\top y^*.$$

For now we only notice some intuition why this may be the case. Provided the false linkage rate is low, $(1-\lambda)/N^* \approx (1-\lambda)/N_2 \approx 0$ for large N^*, and the mis-specification of p_{ij}, where $i \neq j$, may not matter much for the records D_M. Moreover, the proportion of unmatched but linked entities is then also low, so that there are relatively few rows like that for Barkes here. In short, the effects due to the misspecification of the P-matrix may be limited given *low* false linkage rate, and the linkage-model estimators $\widehat{\beta}_{LL}$ and $\widehat{\beta}_A$ may still remove most of the bias of the face-value estimator $\widehat{\beta}^*$.

1.2.2 Non-informative linkage error assumption

The linkage model essentially requires one to specify, for any given record i in D_1, the probability of $a_{ij} = 1$ for *all* the records $j \in D_2$. To accommodate incomplete match space and heterogeneous linkage errors, we specify the *non-informative linkage error (NILE)* assumption as follows:

$$\lambda_i = \Pr(a_{ii} = 1|\ell_i = 1, X, y) = \begin{cases} \Pr(a_{ii} = 1|\ell_i = 1) & \text{for } i \in D_M \\ 0 & \text{for } i \notin D_M \end{cases} \tag{1.2}$$

and, for $i \in D_1$ (or D_2), the probability of linkage is independent of (X, y), i.e.

$$\psi_i = \Pr(\ell_i = 1|X, y) = \Pr(\ell_i = 1). \tag{1.3}$$

Heterogeneous linkage error is the case if λ_i varies over D_M and ψ_i over D_1 (or D_2). The assumption (1.2) accommodates incomplete match space, assigning zero chance of correct link to any unmatched entities in $D_1 \setminus D_M$ or $D_2 \setminus D_M$, without needing to specify p_{ij} for $i \in D_1$, $j \in D_2$ and $j \neq i$. It is possible to incorporate in ψ_i a sample inclusion probability, as when D_1 is a sample from population D_2.

We introduce also a slightly weaker NILE assumption as follows, which we use for the consistency results later on. Let z_i be a well-defined function of x_i and y_i, such as $x_i y_i$ for $i \in D_M$ or $z_i = x_i x_i^\top$ for $i \in D_1$, where D_z is the corresponding entity set of z_i, which is of the size N_z. Let $\bar{\psi} = \sum_{i \in D_z} \psi_i / N_z$, $\bar{z} = \sum_{i \in D_z} z_i / N_z$, and $S_{\psi z} = \sum_{i \in D_z} (\psi_i - \bar{\psi})(z_i - \bar{z}) / N_z$. *Asymptotic NILE over D_z* is the case, as $N_z = |D_z| \to \infty$, provided (1.2) and

$$S_{\psi z} \to 0 \,, \tag{1.4}$$

i.e. ψ_i and z_i are empirically uncorrelated over the set D_z. Notice that (X, y) can be treated as constants in (1.4), to be incorporated in a design-based approach to sample survey data, where record linkage is needed. The assumption (1.4) is weaker than (1.3), since (1.3) implies (1.4), but not *vice versa*.

Since regression analysis is conditional on X, other authors using the linkage-model approach assume non-informative linkage error is the case if a_{ij}'s are independent of y conditional on X (Lahiri & Larsen (2005); Chambers (2009)). While the formulation is parsimonious, in reality it is not weaker than the definition here, as we discuss below. Let c_i be the linkage key variables, and $c_i^{(1)}$ the observed value of c_i in D_1 and $c_i^{(2)}$ that in D_2. In many applications, c_i is not involved in the regression, such as when c_i consists of Name, Date of Birth and Address. It seems reasonable to assume that the potential measurement errors affecting $(c_i^{(1)}, c_i^{(2)})$ are independent of (x_i, y_i) given c_i. Let $C, C^{(1)}, C^{(2)}$ be the matrix notation for $c_i, c_i^{(1)}, c_i^{(2)}$, we would then have

$$\Pr(a_{ii} = 1, \ell_i = 1|X, y, C, C^{(1)}, C^{(2)}) = \Pr(a_{ii} = 1, \ell_i = 1|C, C^{(1)}, C^{(2)}),$$

so that (λ_i, ψ_i) neither depend on y nor X, either conditional on $(C, C^{(1)}, C^{(2)})$ or after integrating out $(C, C^{(1)}, C^{(2)})$ with respect to whichever distribution they have.

It is still possible sometimes that a key variable, which necessarily is present in *both* datasets, may be related to the x-variables, but not the y-variable. For example, Age or Country of Birth may form part of x_i, possibly after some regrouping. Let x_{ic} contain these common variables between c_i and x_i. Let x_i^R be the remaining x-variables, and c_i^R the remaining key variables. The NILE assumption is satisfied provided x_{ic} is used as blocking variables in record linkage, such that only records within the same block can possibly be linked to each other, because the blocking variables are considered to be free of measurement errors. This is typically the case with the variables Age and Country of Birth.

However, it is conceivable that the overlapping x_{ic} is not used as a blocking variable. It is currently an open question how to deal with informative linkage errors. The problem is complicated not least when the observed values $(x_{ic}^{(1)}, x_{ic}^{(2)})$ may differ from the true x_{ic} and, depending on the method of record linkage, $x_{ic}^{(1)}$ may or may not be equal to $x_{ic}^{(2)}$ given $\ell_i = 1$. Thus, the value of x_{ic} to be used in the linkage-data linear regression may be subjected to measurement error, whether or not record i is correctly linked. In this Chapter we shall assume that the potential linkage error due to such key-variable covariates is negligible compared to the rest key variables c_i^R, so that the NILE assumption remains acceptable. The same is needed when non-informativeness is defined conditionally given X.

1.2.3 OLS based on Gold linkage

For the first estimator of β to be considered, we assume the linked set is such that missing match is possible but not false links, to be referred to as a *Gold* linkage procedure. Denote by $D_G^* = D_1^* = D_2^*$ the Gold linkage set, which involves a further selection from all the links that otherwise might have been considered acceptable. Linkage procedures that allow false links are referred to as *sub-Gold* linkage. The terms Gold and sub-Gold are only used as shorthands of the two record linkage settings, and no emotive connotation is intended. We have $\lambda_i = \Pr(a_{ii} = 1|\ell_i = 1) = 1$ by Gold linkage. Denote by $\tilde{\beta}$ the ideal OLS based on $(x,y)_M$, and by $\widehat{\beta}_G$ the OLS based on $(x,y)_{D_G^*}$, which are, respectively,

$$\tilde{\beta} = (\sum_{i \in D_M} x_i x_i^\top)^{-1} (\sum_{i \in D_M} x_i y_i),$$
$$\widehat{\beta}_G = (\sum_{i \in D_G^*} x_i x_i^\top)^{-1} (\sum_{i \in D_G^*} x_i y_i^*) = (\sum_{i \in D_G^*} x_i x_i^\top)^{-1} (\sum_{i \in D_G^*} x_i y_i). \tag{1.5}$$

Proposition 1 Asymptotically, as $N_M = |M| \to \infty$, we have $\widehat{\beta}_G - \tilde{\beta} \xrightarrow{P} 0$, provided

(g1) NILE assumption (1.2), with $\lambda_i \equiv 1$, and (1.4) over D_M,

(g2) $E(N_G^*/N_M) \to \psi > 0$, where $N_G^* = |D_{G^*}|$.

Under the regression model, the variance of $\widehat{\beta}_G$ conditional on $X_{D_G^*}$ is given by

$$V(\widehat{\beta}_G^*) = (X_{D_G^*}^\top X_{D_G^*})^{-1} (X_{D_G^*}^\top V(y_{D_G^*}) X_{D_G^*}) (X_{D_G^*}^\top X_{D_G^*})^{-1}.$$

The convergence can be established directly under (g1) and (g2), where the x- and y-values are treated as constants. For any $z_i = z(x_i, y_i)$ for $i \in D_M$, we have

$$E(\sum_{i \in D_G^*} z_i | X, y) = E(\sum_{i \in D_M} \ell_i z_i | X, y) = \sum_{i \in D_M} E(\ell_i | X, y) z_i = \sum_{i \in D_M} \psi_i z_i$$

Thus, ℓ_i being an unbiased estimator of ψ_i, we have $\sum_{D_G^*} z_i / N_M - \bar{\psi}\bar{z} \xrightarrow{P} 0$, provided (g1). Provided (g2), so that $\bar{\psi} \to \psi$, we have $\sum_{D_G^*} z_i / N_G^* - \bar{\psi}\bar{z} \xrightarrow{P} 0$. The result $\widehat{\beta}_G - \tilde{\beta} \xrightarrow{P} 0$ follows from replacing z_i with $x_i x_i^\top$ and $x_i y_i$ in both the estimators.

We notice that the consistency of $\widehat{\beta}_G$ given by (1.5) holds when record linkage follows sampling from D_1 or D_2 or both, provided sampling is non-informative of the x- and y-values in D_M. Finally, in the case of $V(y_{D_M}) = \sigma^2 I_{N_M \times N_M}$, $V(\widehat{\beta}_G^*)$ reduces to $(X_{D_G^*}^\top X_{D_G^*})^{-1} \sigma^2$. The relative efficiency to the ideal $\tilde{\beta}$ converges to $1/\psi$, as $N_M \to \infty$ asymptotically.

1.2.4 Covariance of (x_i, y_i^*)

To estimate β based on sub-Gold linkage, we shall make use of the covariance between x_i and its linked y-value. The result below holds for any analysis of interest, not just linear regression. For any $i \in D_1$, we observe x_i. At most one link is allowed for each record. For any linked record $i \in D_1^*$, its linked y-value is given by $y_i^* = \sum_{j \in D_2} a_{ij} y_j$. Provided NILE

(1.2), for any $i \in D_M$, we have

$$\begin{aligned} Cov(x_i, y_i^*|\ell_i = 1) &= Cov(x_i, a_{ii}y_i|\ell_i = 1) + \sum_{j\neq i} Cov(x_i, a_{ij}y_j|\ell_i = 1) \\ &= E(a_{ii}|\ell_i = 1)Cov(x_i, y_i) + \sum_{j\neq i} E(a_{ij}|\ell_i = 1)Cov(x_i, y_j). \end{aligned}$$

As long as x_i and y_j are uncorrelated for $i \neq j$, we have $Cov(x_i, y_i^*|\ell_i = 1, a_{ii} = 1) = Cov(x_i, y_i)$ given correct linkage, and $Cov(x_i, y_i^*|\ell_i = 1, a_{ii} = 0) = 0$ given false link of any matched entity $i \in D_M$, or linkage of an unmatched unit $i \in D_1 \setminus D_M$. It follows that, for any $i \in D_1$,

$$Cov(x_i, y_i^*|\ell_i = 1) = \lambda_i Cov(x_i, y_i),$$

where λ_i is given by (1.2). That is, false links on average move the observed covariance among the linked pairs of records towards zero. Moreover, to account for the effective *matched* sample size of the empirical covariance between x_i and y_i^* over the linked set D_1^*, one only needs to know the total number of correct matches in D_1^*, but not necessarily the individual λ_i's. The idea is developed below.

1.2.5 Pseudo-OLS based on sub-Gold linkage

Given any sub-Gold linkage procedure, let the Pseudo-OLS fit of β be given by

$$\begin{aligned} \widehat{\beta}_P &= (\frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*})^{-1} (\bar{x}\bar{y}^* + \widehat{\lambda}^{-1} S_{xy*}) & (1.6) \\ &= \widehat{\lambda}^{-1}\widehat{\beta}^* - (\widehat{\lambda}^{-1} - 1)(\frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*})^{-1} \bar{x}\bar{y}^*, & (1.7) \end{aligned}$$

where $\bar{x} = \sum_{i\in D_1^*} x_i/N^*$, and $\bar{y}^* = \sum_{i\in D_1^*} y_i^*/N^*$, and $S_{xy^*} = \sum_{i\in D_1^*}(x_i - \bar{x})(y_i^* - \bar{y}^*)/N^*$, and $\widehat{\lambda}$ is an estimate of the number of correct matches among the actual links. Notice that $\widehat{\lambda}$ can be obtained for the realised D_1^*. The expression (1.6) reveals that the Pseudo-OLS is based on a linkage-error adjustment of the observed covariance between x_i and y_i^* in the linked dataset, whilst the expression (1.7) shows it as a linear adjustment of the naïve face-value OLS $\widehat{\beta}^* = (X_{D_1^*}^\top X_{D_1^*})^{-1} X_{D_1^*}^\top y^*$.

Example For simple linear regression $y_i = \alpha + \beta x_i + \epsilon_i$, the Pseudo-OLS is given by

$$\begin{bmatrix} \widehat{\alpha}_P \\ \widehat{\beta}_P \end{bmatrix} = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N^*}\sum_{i\in D_1^*} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \bar{y}^* \\ \bar{x}\bar{y}^* + \widehat{\lambda}^{-1} S_{xy^*} \end{bmatrix}$$

$$\Rightarrow \quad \widehat{\beta}_P = \widehat{\lambda}^{-1}\frac{S_{xy^*}}{S_x^2} = \widehat{\lambda}^{-1}\frac{\sum_{i\in D_1^*}(x_i - \bar{x})(y_i^* - \bar{y}^*)}{\sum_{i\in D_1^*}(x_i - \bar{x})^2} \quad \text{and} \quad \widehat{\alpha}_P = \bar{y}^* - \bar{x}\widehat{\beta}_P,$$

where $\widehat{\beta}_P$ is a multiplicative adjustment of the face-value OLS of the slope *away from 0*, for $\widehat{\lambda} < 1$. This is intuitive because, given a false link is made for $i \in D_1^*$, the face-value covariance $(x_i - \bar{x})(y_i^* - \bar{y}^*) = (x_i - \bar{x})(y_j - \bar{y}^*)$ has approximately expectation zero, as long as x_i and y_j are uncorrelated for $j \neq i$. So the face-value estimate of the slope is biased towards 0. To adjust for the bias, notice that the effective sample size underlying the linked sample covariance S_{xy^*} is just the number of true matches among the links, which is estimated by $\widehat{\lambda}N^*$. This is the basic idea underlying the Pseudo-OLS (1.6).

Consistency conditions for Pseudo-OLS

Given sub-Gold linkage, we have $E(N^*_{MM}|D^*_{1M}) = \sum_{i\in D^*_{1M}} \lambda_i = \sum_{i\in D^*_1} \lambda_i = E(N^*_{MM}|D^*_1)$, because $\lambda_i = 0$ for the unmatched entities in $D^*_1 \backslash D^*_{1M}$. We have $\widehat{\beta}_P - \tilde{\beta} \overset{P}{\to} 0$, if the difference between each term in (1.6) and its counterpart in $\tilde{\beta}$ converges to zero in probability. In addition to the NILE assumption and the consistency of $\widehat{\lambda}$, regularity conditions are needed regarding the values of (x, y) associated with the matched entities in D_M and x (or y) of the unmatched entities in D_1 (or D_2). All the conditions are given below in Proposition 2, the proof of which is given in Appendix 1.6.

Proposition 2 Asymptotically, as $N_M = |M| \to \infty$, we have $\widehat{\beta}_P - \tilde{\beta} \overset{P}{\to} 0$, provided

(p0.1) $Cov(x_i, y_j) = 0$ for $j \neq i$, $i \in D_1$ and $j \in D_2$,

(p0.2) $\sum_{i\in D_M} x_i/N_M - \sum_{i\in D_1} x_i/N_1 \to 0$,

(p0.3) $\sum_{j\in D_M} y_j/N_M - \sum_{j\in D_2} y_j/N_2 \to 0$,

(p1) NILE assumption (1.2) and (1.4), where (1.4) holds over D_1 as well as D_2,

(p2) $E(N^*) \to \infty$, and $E(N^*_{MM}/N^*) \to \lambda > 0$, and $\widehat{\lambda} \overset{P}{\to} \lambda$.

Variance estimation

It is impractical to allow heterogeneous variance of ϵ_i, because we do not know the x-values in the case of a false link. We shall therefore assume $V(y_i) = \sigma^2$ for all $i \in D_2$. Provided NILE, it is natural to condition on the realised N^*. Given false link of $i \in D^*_1$, we have $y^*_i = y_j$, for some $j \in D_2$ and $j \neq i$, where the record j may or may not belong to D_M. In the case of $j \notin D_M$, we shall assume that there nevertheless exists a vector x_j under the regression model, even though $j \notin D_1$. Thus, we shall condition on $(X_{D_1}, X_{D_2\backslash D_M}, N^*)$ throughout the following. We have

$$V(\widehat{\beta}_P) = (\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1})^{-1} V(\bar{x}\bar{y}^* + \widehat{\lambda}^{-1} S_{xy^*})(\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1})^{-1}.$$

Now, given the linkage matrix $A = [a_{ij}]$, where at most one link is allowed for a record, $y^*_i = y_j$ is conditionally independent of $y^*_k = y_l$ for $i \neq k$, since $j \neq l$ regardless if (i, j) and (k, l) are true matches or not. Thus, we have

$$V(\bar{x}\bar{y}^* + \widehat{\lambda}^{-1} S_{xy^*}) = V(\bar{x}\bar{y}^*) + V(\widehat{\lambda}^{-1} S_{xy^*}),$$

since $Cov(\bar{y}^*, y^*_i - \bar{y}^*|A) = 0$, hence $Cov(\bar{y}^*, \widehat{\lambda}^{-1} S_{xy^*}|A) = 0$ and $Cov(\bar{y}^*, \widehat{\lambda}^{-1} S_{xy^*}) = 0$. By working out $V(\bar{x}\bar{y}^*)$ and $V(\widehat{\lambda}^{-1} S_{xy^*})$ – see Appendix 1.6 for details, we obtain

$$V(\widehat{\beta}_P) \approx (X^\top_{D^*_1} X_{D^*_1})^{-1}\sigma^2 + (\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1})^{-1} \Delta (\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1})^{-1}, \qquad (1.8)$$

where

$$S_{xx} = \frac{1}{N^*} \sum_{i\in D^*_1} (x_i - \bar{x})(x_i - \bar{x})^\top \quad \text{and} \quad \Delta = (\frac{1}{\lambda^2} - 1)\frac{\sigma^2}{N^*} S_{xx} + V(\widehat{\lambda}) S_{xx}\beta\beta^\top S^\top_{xx}.$$

Clearly, linkage errors cause a loss of efficiency, since the first term on the right-hand side of (1.8) would have been the variance had all the links been true matches and adjustment not needed. The extra variance depends on Δ, which has two contributing terms: one due to the smaller effective sample size N^*_{MM} compared to the face-value sample size N^*, the other due to the estimation uncertainty of the adjustment factor $\widehat{\lambda}$. Compared to $\widehat{\beta}_G$ by Gold linkage, the first term of (1.8) is smaller than $V(\widehat{\beta}_G)$, since $D^*_G \subset D^*_1$. However, the extra uncertainty in (1.8) due to Δ may still possibly cause loss of efficiency of sub-Gold linkage compared to Gold linkage. The matter is explored empirically in Section 1.5.

For plug-in variance estimation, we need an estimate of σ^2, in addition to $\widehat{\beta}_P$ and $\widehat{\lambda}$. Applying the standard formula of OLS variance estimator to the linkage data, we obtain

$$S^*_{ee} = \frac{1}{N^* - p} \sum_{i \in D^*_1} (y^*_i - \widehat{\beta}^\top_P x_i)^2 = \frac{1}{N^* - p} \sum_{i \in D^*_1} [(y_{j_i} - \widehat{\beta}^\top_P x_{j_i}) - \widehat{\beta}^\top_P (x_i - x_{j_i})]^2,$$

$$E(S^*_{ee}) \xrightarrow{P} \sigma^2 + 2(1-\lambda)\beta^\top E(S_{xx})\beta,$$

as $N_M \to \infty$, where $j_i \in D_2$ is linked to $i \in D_1$, and $(x_i - x_{j_i}) = 0$ with probability λ_i. The face-value estimator of σ^2 has therefore an upwards bias asymptotically, which is bounded by the overall false linkage rate $1 - \lambda$, and can be adjusted accordingly.

1.2.6 Asymptotic bias when using the ELE-model

The ELE-model treats incomplete match space as if it were complete. To examine the resulting bias, consider $\widehat{\beta}_A = \left(X^\top_{D^*_1} P(\lambda) X_{D^*_1}\right)^{-1} X^\top_{D^*_1} y^*$, where

$$P(\lambda) = \lambda I_{N^* \times N^*} + \lambda_{N^*}(\mathbf{1}\mathbf{1}^\top - I)_{N^* \times N^*}, \qquad \lambda_{N^*} = \frac{1-\lambda}{N^* - 1},$$

$$X^\top_{D^*_1} P(\lambda) X_{D^*_1} = G + H, \quad G = \lambda X^\top_{D^*_1} X_{D^*_1}, \quad H = \lambda_{N^*} N^* (N^* \bar{x}\bar{x}^\top - \frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1}).$$

An estimate of the *overall* true match rate among the links is used as $\widehat{\lambda}$. By a Lemma due to Miller (1981): $(G+H)^{-1} = G^{-1} + (1+g)^{-1} G^{-1} H G^{-1}$, where $g = \text{tr}(HG^{-1})$, we can write

$$\widehat{\beta}_A(\lambda) = \frac{1}{\lambda}\widehat{\beta}^* - \frac{\lambda_{N^*} N}{\lambda^2 (1+g)} (X^\top_{D^*_1} X_{D^*_1})^{-1} (N^* \bar{x}\bar{x}^\top - \frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1}) \widehat{\beta}^*.$$

Let $\bar{x}^\top (\frac{1}{N} X^\top_{D^*_1} X_{D^*_1})^{-1} \bar{x} \xrightarrow{P} \kappa_x$, as $N^* \to \infty$, we have

$$\begin{aligned} g &= \text{tr}\left(\frac{\lambda_{N^*} N^*}{\lambda} (N^* \bar{x}\bar{x}^\top - \frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1})(X^\top_{D^*_1} X_{D^*_1})^{-1}\right) \\ &= \frac{\lambda_{N^*} N^*}{\lambda} (\bar{x}^\top (\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1})^{-1} \bar{x} - \frac{p}{N^*}) \quad \xrightarrow{P} \quad \frac{1-\lambda}{\lambda} \kappa_x. \end{aligned}$$

Let $(\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1})^{-1} \bar{x}\bar{x}^\top \to \zeta$, as $N^* \to \infty$. Provided consistent Pseudo-OLS, we have

$$\widehat{\beta}_A - \widehat{\beta}_P \quad \xrightarrow{P} \quad \frac{1-\lambda}{\lambda} \zeta\beta - \frac{1-\lambda}{\lambda(\kappa_x + (1-\kappa_x)\lambda)} \zeta(\beta + E(\widehat{\beta}^* - \beta)),$$

which is the asymptotic bias of $\widehat{\beta}_A$. In cases $\kappa_x \approx 1$ and $\lambda \approx 1$, the asymptotic bias is of the magnitude $(1-\lambda)\zeta E(\widehat{\beta}^* - \beta)$, which is bounded by the false linkage rate $1 - \lambda$. Then, direct application of the ELE-model estimator can nevertheless remove almost all the bias of the face-value OLS.

Example Consider $y_i = \alpha + \beta x_i + \epsilon_i$. Let $\sum_{i \in D_1^*} x_i^*/N^* \xrightarrow{P} \mu_x$ and $\sum_{i \in D_1^*} (x_i^*)^2/N^* \xrightarrow{P} \tau_x$. We have

$$\kappa_x = [1 \;\; \mu_x] \begin{bmatrix} 1 & \mu_x \\ \mu_x & \tau_x \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \mu_x \end{bmatrix} = \frac{\tau_x - \mu_x^2}{\tau_x - \mu_x^2} = 1.$$

It follows that the asymptotic bias of $\widehat{\beta}_A$ based on the ELE-model is given by

$$E(\widehat{\beta}_A - \beta) = -\frac{1-\lambda}{\lambda}\zeta E(\widehat{\beta}^* - \beta) = -\frac{1-\lambda}{\lambda}\begin{bmatrix} [1 \;\; \mu_x]\, E(\widehat{\beta}^* - \beta) \\ 0 \end{bmatrix},$$

$$\zeta = \begin{bmatrix} 1 & \mu_x \\ \mu_x & \tau_x \end{bmatrix}^{-1} \begin{bmatrix} 1 & \mu_x \\ \mu_x & \mu_x^2 \end{bmatrix} = \begin{bmatrix} 1 & \mu_x \\ 0 & 0 \end{bmatrix}.$$

In other words, the slope estimator is unbiased asymptotically, as $N^* \to \infty$; the bias of the intercept estimator is negligible as well, e.g. it is about 2% of the bias of the face-value OLS if the overall false linkage rate is 2%, despite heterogeneous linkage errors. This is thus a favourable setting, under which the estimator can be robust against departures from ELE-model assumptions.

1.3 A diagnostic test for NILE

In one form or another, assumptions of non-informative linkage errors are required in all the existing least-squares methods. For $\widehat{\beta}_G$ and $\widehat{\beta}_P$ developed above, it is natural to ask if one can test whether the NILE assumption is acceptable in a given application. Provided both the estimators are consistent, we have $\widehat{\beta}_G - \widehat{\beta}_P \xrightarrow{P} 0$, as $N_G^* \to \infty$, which suggests the following diagnostic test statistic

$$t = (\widehat{\beta}_G - \widehat{\beta}_P)^\top V(\widehat{\beta}_G - \widehat{\beta}_P)^{-1} (\widehat{\beta}_G - \widehat{\beta}_P) \sim \chi_p^2 \tag{1.9}$$

for H_0 : NILE (g1) and (p1) vs. H_1 : not both (g1) and (p1). Provided asymptotic normal distribution of $\widehat{\beta}_G - \widehat{\beta}_P$, as $N_M = |M| \to \infty$, t follows the χ_p^2-distribution. The test (1.9) bears some resemblance to that of Hausman (1978). However, neither $\widehat{\beta}_G$ nor $\widehat{\beta}_P$ is consistent under H_1, and neither of them is fully efficient under H_0. In addition, $\widehat{\beta}_P$ also involves the estimate of the parameter λ in the denominator, thus, the power of the test can be limited compared to that of Hausman (1978), and we need to derive the variance $V(\widehat{\beta}_G - \widehat{\beta}_P)$ directly.

Let D_G^* and D_P^* be the set of linked entities from D_1 under Gold and sub-Gold linkage, respectively. The variance $V(\widehat{\beta}_P)$ is given by (1.8) on replacing D_1^* with D_P^*, whereas $V(\widehat{\beta}_G) = (X_{D_G^*}^\top X_{D_G^*})^{-1}\sigma^2$. As shown in Appendix 1.6, the covariance $Cov(\widehat{\beta}_G, \widehat{\beta}_P)$ can be given by

$$\begin{aligned} Cov(\widehat{\beta}_G, \widehat{\beta}_P) &\approx \frac{\sigma^2}{\lambda N_P^*} H_G(\bar{x}_G \bar{x}_G^\top + S_G^2) H_P + (1 - \frac{1}{\lambda}) \frac{\sigma^2}{N_P^*} H_G \bar{x}_G \bar{x}_P^\top H_P \\ &= \frac{\sigma^2}{\lambda N_P^*} H_P - (\frac{1}{\lambda} - 1) \frac{\sigma^2}{N_P^*} H_G \bar{x}_G \bar{x}_P^\top H_P, \end{aligned} \tag{1.10}$$

where $H_G = (\frac{1}{N_G^*}\sum_{i \in D_G^*} x_i x_i^\top)^{-1}$ and $\bar{x}_G = \frac{1}{N_G^*}\sum_{i \in D_G^*} x_i$, and $H_P = (\frac{1}{N_P^*}\sum_{i \in D_P^*} x_i x_i^\top)^{-1}$ and $\bar{x}_P = \frac{1}{N_P^*}\sum_{i \in D_P^*} x_i$. Notice that, in case $\lambda \approx 1$, the covariance is dominated by the first term, and the difference between the first terms of (1.10) and (1.8) is positive definite since

$1/\lambda > 1$. Moreover, λN_P^* is the asymptotic expectation of the number of true matches by sub-Gold linkage, which can easily be larger than N_G^* unless all the additional links are false. One may therefore expect positive definite $V(\widehat{\beta}_G) - Cov(\widehat{\beta}_G, \widehat{\beta}_P)$, since $H_P - H_G \overset{P}{\to} 0$ provided the consistency conditions for $\widehat{\beta}_G$ and $\widehat{\beta}_P$.

1.4 An application to income data

The data of this application refers to administrative tax registers of income declarations in 2014 and 2015. The linkage procedure aims to connect incomes in the two years related to the same individuals. The linkage key variables are generally of good quality, though in some cases they can be missing or affected by errors. The linkage is carried out at the Italian National Institute of Statistics, and the false linkage rate is assessed to be between 1.18% and 3.76%. No information about the linkage errors at the individual level are available to us. We consider a simple linear regression model, where the income in 2014 is treated as x and that in 2015 as y. The analysis here is concerned with the data from a small locality, where there are 791 individuals in the tax register in 2014 and 771 in 2015. The linked set contains 711 individuals. A scatter plot of the associated $(x, y)_{M^*}$ is given in Figure 1.2. The application illustrates an advantage of using administrative data, which allows one to carry out analysis at a detailed level that cannot be supported by sample surveys otherwise.

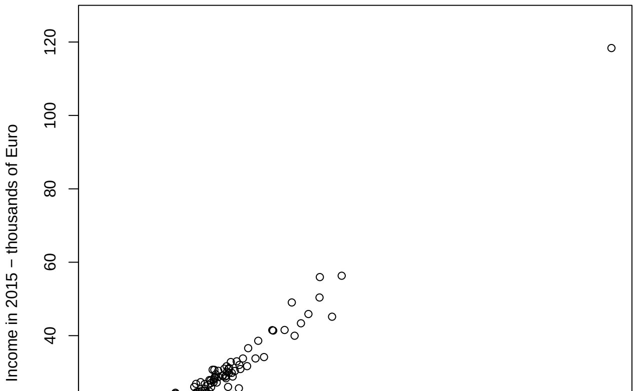


Figure 1.2. Scatter plot of linked income data in the application.

For this linkage dataset, we calculate the face-value OLS $\widehat{\beta}^*$, the estimators $\widehat{\beta}_{LL}$ and $\widehat{\beta}_A$ under the ELE-model, as well as the Pseudo-OLS $\widehat{\beta}_P$ that allows for heterogeneous linkage errors and incomplete match space. Without information about the false linkage probabilities of the individual links, we cannot further select a Gold linkage set D_G^*, or implement the diagnostic test (1.9). The Gold-linkage OLS $\widehat{\beta}_G$ and the diagnostic test (1.9) will be investigated in a simulation study in Section 1.5.

Table 1.1 shows the estimated regression coefficients and their associated confidence

Table 1.1. Estimates of year-on-year income intercept and slope, with associated confidence intervals

	False linkage rate fixed at 1.18%			
Estimator	Intercept	Confidence Interval	Slope	Confidence Interval
$\widehat{\beta}^*$	90.644	[-114.217 , 295.505]	0.983	[0.968 , 0.998]
$\widehat{\beta}_{LL}$	7.191	[-242.640 , 257.023]	0.994	[0.961 , 1.028]
$\widehat{\beta}_A$	52.242	[-139.454 , 243.938]	0.983	[0.964 , 1.002]
$\widehat{\beta}_P$	7.310	[-129.794 , 144.414]	0.994	[0.984 , 1.005]
	False linkage rate fixed at 3.76%			
Estimator	Intercept	Confidence Interval	Slope	Confidence Interval
$\widehat{\beta}^*$	90.644	[-114.217 , 295.505]	0.983	[0.968 , 0.998]
$\widehat{\beta}_{LL}$	-182.411	[-598.275 , 233.451]	1.021	[0.960 , 1.082]
$\widehat{\beta}_A$	-125.782	[-407.104 , 155.541]	1.007	[0.976 , 1.037]
$\widehat{\beta}_P$	-182.012	[-330.779 , -33.246]	1.021	[1.001 , 1.032]

intervals. The face-value OLS suggests that the regression model can explain most of the variation in the dependent variable ($R^2 = 0.958$). In particular, the relative standard error of the slope estimator is only 0.007, which is of the same magnitude as the aforementioned false linkage rates. It follows that the bias due to the false links is not a negligible source of error, compared to the variance of the slope estimator, so that appropriate adjustment of the linkage errors is important in this case.

Fixing the overall false linkage rate $1 - \lambda$ either at 1.18% or 3.76%, the other estimates and their associated confidence intervals are given in Table 1.1. It can be seen that $\widehat{\beta}_A$ deviates least from the face-value OLS, for both values of λ; whereas $\widehat{\beta}_{LL}$ and $\widehat{\beta}_P$ are close to each other. However, the Pseudo OLS $\widehat{\beta}_P$ is apparently much more efficient compared to the ELE-model estimators. For example, at $1 - \lambda = 1.18\%$, the width of the confidence interval is 0.067 for the slope estimator by $\widehat{\beta}_{LL}$, whereas it is 0.021 by $\widehat{\beta}_P$, according to which the variance ratio between the two is only about 10%. The efficiency gain is somewhat greater at $1 - \lambda = 3.76\%$.

A reason that the Pseudo-OLS can be more efficient than the ELE-model estimators is that the linkage-error adjustment affects only the linked sample covariance S_{xy^*}, but not the marginal sample quantities such as the means of x and y or the matrix $X_{D_1^*}^\top X_{D_1^*}$. Of course, there is the possibility that the comparison here may be affected by the quality of variance estimation, so that the relative efficiency is not accurately assessed. We shall examine this point in the simulation study in Section 1.5.

The variance formula (1.8) allows one to incorporate the estimation uncertainty in $\widehat{\lambda}$, which is not available to the existing ELE-model estimators in closed-form expression. Since we are not provided an estimate of $V(\widehat{\lambda})$, we proceed in a practical manner as follows. Treating the reported range of false linkage rate as if it were a 95% normality-based confidence interval for $1 - \lambda$, we obtain the centre point $1 - \widehat{\lambda} = 2.47\%$ as an estimate of $1 - \lambda$, and we use the quarter length 0.645% as an estimate of $SE(\widehat{\lambda})$. Applying $\widehat{\beta}_P$ with this $\widehat{\lambda}$ and its associated estimate of $V(\widehat{\lambda})$, we obtain the regression coefficient estimates -86.099 and 1.008 for the intercept and slope, respectively, with associated confidence interval [-258.478 , 86.279] for the intercept and [0.991, 1.024] for the slope. As can be expected, the point estimates are between the corresponding ones reported in Table 1.1. The width of the confidence interval for the slope is now 0.033, compared to 0.031 when $1 - \lambda$ is fixed at 3.76% and 0.021 when $1 - \lambda$ is fixed at 1.18%. Thus, it would be misleading if the inference does not take into account the uncertainty due to the estimation of λ. This is another advantage of the Pseudo-OLS method.

In this application, an interesting development is the linkage adjustment of two-part models for semi–continuous data, which may be appropriate to deal with concentration of zeros in income values.

1.5 A simulation study

We have four main objectives for this simulation study. First, we would like to be reassured that the apparent efficiency gains of the proposed Pseudo-OLS is not misleading. Second, a related question is the quality of associated variance estimation. Third, since one does not know to what extent the assumption of exchangeable linkage errors is violated in the application, confirmation can be obtained by simulation that the Pseudo-OLS estimator does hold in the presence of heterogeneous linkage errors. Four, we would like to investigate the performance of the diagnostic test for the NILE assumption. To the end of these objectives, we devise three scenarios below in Section 1.5.1.

1.5.1 Set-up

Scenario I: Real-life linkage and regression data. This scenario addresses all the four objectives.

The ESSnet-DI is a Eurostat project on data integration from 2009 to 2011. We use the data disseminated by ESSnet-DI (Heasman & (2011)), which are freely available online. The dataset comprises over 26000 individuals. It contains synthetic linkage key variables (names, dates of birth, addresses) for each individual. The key variables are distorted by missing values and typos in several different ways, which imitate real-life errors in these variables that can cause potential linkage errors. One can observe the true linkage errors by comparing the links with the true matches that are known.

For real-life regression data, we attach anonymised income data to each individual in the ESSnet-DI population, which are drawn randomly and with replacement from the linked tax data, but without being limited to the locality (in Section 1.4) with only 711 linked records. A scatter plot of the synthetic population income data is given in Figure 1.3. It can be seen that a simple linear regression model remains plausible for the simulated population values. However, there are now clearly outliers to the regression model, drawn from outside the data in the application (Figure 1.2). We do not remove the outliers, since it would be interesting to explore how they might affect the results.

To simulate repeated linkage and regression analysis, each time we draw first a sample of 1000 individuals from this fixed synthetic population. We then break up the sample into two separate sets D_1 and D_2, where D_1 is selected from the 1000 individuals by Bernoulli sampling with probability $\pi_1 = 0.93$, and D_2 by separate Bernoulli sampling with probability $\pi_2 = 0.92$. This creates an incomplete match space, where the expected number of matched individuals between D_1 and D_2 is $1000\pi_1\pi_2 \approx 856$.

Using a chosen set of key variables, probabilistic linkage by the approach of Fellegi & Sunter (1969) is implemented using the software Relais (2015). Over 100 simulations, the average match rate N^*_{MM}/N_M is 83.3% and the false linkage rate $1 - N^*_{MM}/N^*$ is 2.016%, i.e., the sub-Gold linkage setting. For Gold linkage, we use a different set of key variables with fewer errors. Over 100 simulations, the average match rate is reduced to about 50%, while the false linkage is reduced to 0.046%. The linkage errors are heterogeneous across the different individuals.

We apply $\widehat{\beta}_G$ by (1.5) to each Gold linkage set. For each sub-Gold linkage set, we obtain $\widehat{\beta}_P$ by (1.6), as well as the ELE-model estimators $\widehat{\beta}_{LL}$ and $\widehat{\beta}_A$. For these adjustments we use the true overall false linkage rate λ in each linked set. We do not simulate additional

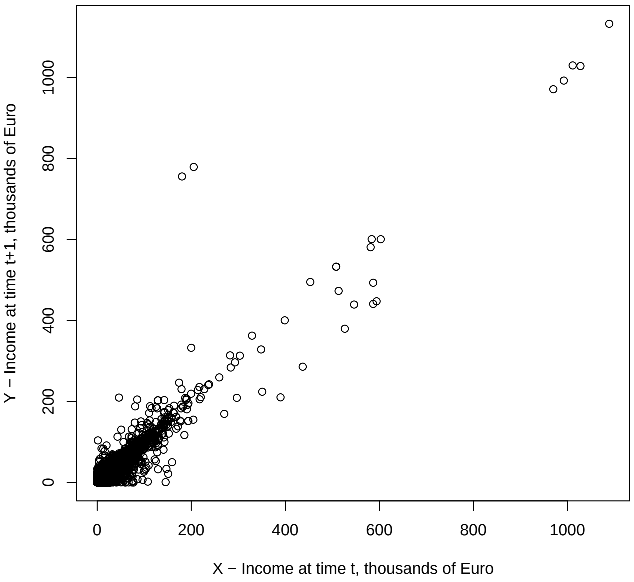


Figure 1.3. Scatter plot of synthetic income data in the ESSnet-DI population.

estimation of λ, as it is not in the focus of this Chapter and it would affect all the adjustment methods equally. Finally, we apply the diagnostic test (1.9) based on $\widehat{\beta}_G$ and $\widehat{\beta}_P$.

Scenario II: Real-life linkage data, artificial regression data. We expect Scenario-I can help us to better understand the application results in Section 1.4. Insofar as the fixed synthetic population of income data may have certain peculiar features that complicate the interpretation, we generate additional artificial regression data, reusing the simple linear regression setting of Chambers (2009), where

$$y_i = 1 + 5x_i + \epsilon_i, \qquad x_i \sim \text{Uniform}(0,1) \qquad \text{and} \qquad \epsilon_i \sim N(0,1).$$

Since the linear regression model holds, while the linkage errors remain uncontrolled and realistic, in Scenario-II we are able to isolate the effects of linkage errors on the estimators. The simulation of repeated linkage and regression analysis is the same as under Scenario-I, except that for each sample of 1000 individuals, we now simulate (x_i, y_i), for $i = 1, ..., 1000$, independently according to the specific regression model above. Regression analysis is then based on these (x, y)-values instead of the real-life income data.

Scenario III: Artificial linkage and regression data. To confirm that $\widehat{\beta}_G$ and $\widehat{\beta}_P$ can deal with heterogeneous linkage errors under the NILE assumption, we simulate artificial linkage data by reusing the setting of Chambers (2009). For each sample of 1000 individuals, we first simulate artificial (x, y)-values as in Scenario-II. Next, the 1000 individuals are randomly divided into three blocks. The first block contains 75% of the individuals, where $\lambda_i \equiv 1$, so that these can be linked perfectly. The second block contains 15% individuals, where $\lambda_i \equiv 0.95$, so that the linkage results would be fairly good for them. The third block contains the remaining 10% individuals, where $\lambda_i \equiv 0.75$, and the linkage results would be

Table 1.2. Results for variance estimation over 100 simulations

		Intercept					
Scenario		True	Naïve	$\widehat{\beta}_{LL}$	$\widehat{\beta}_A$	$\widehat{\beta}_P$	$\widehat{\beta}_G$
I	Standard Error	386.1	457.1	1222.7	575.8	388.9	545.1
	SE Estimator	2431.7	2604.5	2645.2	1256.1	2645.2	3233.6
II	Standard Error	0.069	0.077	0.079	0.079	0.075	0.098
	SE Estimator	0.078	0.086	0.086	0.087	0.086	0.098
III	Standard Error	0.043	0.042	0.043	0.044	0.043	0.051
	SE Estimator	0.045	0.048	0.048	0.047	0.046	0.052
		Slope					
Scenario		True	Naïve	$\widehat{\beta}_{LL}$	$\widehat{\beta}_A$	$\widehat{\beta}_P$	$\widehat{\beta}_G$
I	Standard Error	0.012	0.015	0.052	0.022	0.013	0.018
	SE Estimator	0.113	0.118	0.120	0.057	0.120	0.149
II	Standard Error	0.119	0.134	0.138	0.138	0.131	0.171
	SE Estimator	0.131	0.149	0.150	0.151	0.150	0.160
III	Standard Error	0.075	0.074	0.075	0.074	0.077	0.081
	SE Estimator	0.078	0.083	0.084	0.083	0.080	0.089

rather poor for them. Moreover, we do not simulate subsampling of D_1 and D_2, so that we have complete match space by construction. The linked set can now be simulated directly, without actually implementing any linkage procedure. Had we broken up the sample into D_1 and D_2, dividing the 1000 records in three blocks, and linked every record in D_1 to one in D_2 from the same block, the linkage errors would have been on expectation the same as we have just specified.

This yields an overall false linkage rate that equals to 0.9675, which is quite close to that in Scenario-I (and II). Given each simulated linkage set, we calculate $\widehat{\beta}_P$ using a single adjustment factor $\lambda = 0.9675$, and the ELE-model estimators given block-diagonal P-matrix with known λ-values. We can see how well $\widehat{\beta}_P$ handles heterogenous linkage errors by comparing it to the benchmark ELE-model estimators. Finally, we simply calculate $\widehat{\beta}_G$ based on the first-block of links.

1.5.2 Results of regression coefficient estimators

Figure 1.4 shows the Percentage Relative Errors (PREs) of the different regression coefficient estimates. For each linked set, the 'error' of an estimate is calculated as its difference to the corresponding true OLS estimate $\widehat{\beta}$, based on the matched individuals D_M as when linkage is unnecessary. Over the top margin of each box-plot, we report the actual coverage rates of the nominal 95% confidence intervals using the associated variance estimators. Table 1.2 provides the empirical standard error (SE) of each estimator over the 100 simulations, and the corresponding average of the 100 SE estimates.

Consider the results under Scenario-I, which are immediately relevant to those in Section 1.4. First, as expected, the presence of false links weakens the observed correlation between x_i and y_i^*. Hence, the face-value estimate of the slope is negatively biased when the true slope is positive, and the intercept estimate is biased in the opposite direction. It can be seen in Figure 1.4 that all the adjusted estimators are less biased than the face-value OLS, where $\widehat{\beta}_{LL}$ and $\widehat{\beta}_P$ have the most similar expectations, which is compatible with the application results in Table 1.1, where these two estimators are closest to each other. Moreover, it illustrates that in case of heterogeneous but low false linkage probabilities, the ELE-model estimators can nevertheless remove most of the bias, as discussed in Section 1.2.6.

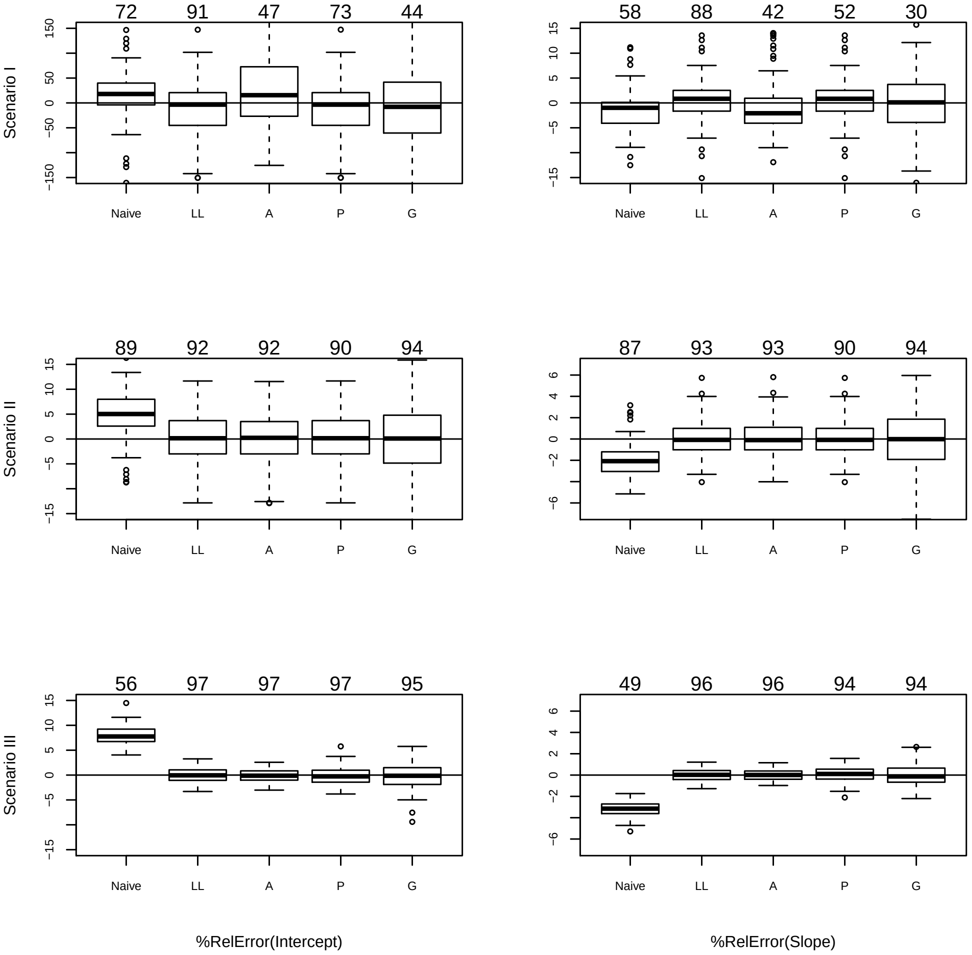


Figure 1.4. Boxplot of PREs of intercept and slope estimates: Scenario-I (top), II (middle), III (bottom); estimator $\widehat{\beta}_{LL}$ (LL), $\widehat{\beta}_A$ (A), $\widehat{\beta}_P$ (P) and $\widehat{\beta}_G$ (G); coverage of 95% confidence interval (over top margin).

Next, according to the SEs in Table 1.2, the Pseudo-OLS is the most efficient of all the linkage-data estimators, including the face-value OLS. The relative efficiency to the ELE-model estimators is comparable to that estimated in Table 1.1. This suggests that the gains are genuine in the application. Since $\widehat{\beta}_P$ is calculated using λ instead of its estimate here, the efficiency gains against $\widehat{\beta}_G$ are somewhat over-stated. Nevertheless, generally one may expect $\widehat{\beta}_P$ to be more efficient than $\widehat{\beta}_G$, as long as the effective sample size N_{MM}^* is much larger based on sub-Gold linkage (e.g. with about 2% false linkage rate here) than based on Gold linkage (e.g. with about 50% missing match rate here).

Meanwhile, the means of the SE estimators (Table 1.2) over the 100 simulations show that all the variances are over-estimated considerably, including the true OLS, and the coverage of the 95% confidence intervals are very erratic. This is mainly caused by the regression model outliers in this case, as noticed earlier for Figure 1.3. Thus, these results serve well as a reminder that, in linkage-data regression, one must not forget about the problems that can also cause troubles in the absence of linkage errors. Notice that variance over-estimation is not a problem for the application results in Table 1.1, where critical outliers are absent from the linked dataset (Figure 1.2).

When it comes to Scenario-II, we can see in Figure 1.4 that all the adjusted estimators are nearly unbiased, as can be expected given the results under Scenario-I. The Pseudo-OLS remains the most efficient linkage-data method. The results of variance estimation appear acceptable for all the estimators, now that outlier-contaminated income data are replaced by true regression data. While there still exists some slight over-estimation of the variance, it is not related to the adjustment methods, because the amount of over-estimation for them is comparable to that for the true OLS. The coverage of the confidence interval derived from the face-value OLS is improved compared to that in Scenario-I, because its bias is relatively small here. Nevertheless, bias adjustment is preferable.

The ELE-model assumptions of $\widehat{\beta}_{LL}$ and $\widehat{\beta}_A$ are fully satisfied in Scenario-III. Likewise for $\widehat{\beta}_G$ under the NILE assumption. Despite $\widehat{\beta}_P$ uses only an overall false linkage rate, Figure 1.4 shows clearly that it is as effective as the benchmark estimators at reducing the bias due to the linkage errors. This confirms that the Pseudo-OLS can accommodate heterogeneous linkage errors in a simple manner, provided the NILE assumption is satisfied. The Pseudo-OLS is no longer the most efficient method here, which is not surprising given that the assumptions of the other estimators are exactly satisfied. The principal advantages of the Pseudo-OLS lies in real-life situations, where the match space is incomplete and the secondary analyst has no detailed knowledge of the record linkage procedure, such as the three blocks of linkage errors in this case. The results of variance estimation are acceptable for all the estimators. Due to increased bias relative to its variance, the face-value OLS again leads to low coverage here. The coverages rates derived from $(\widehat{\beta}_{LL}, \widehat{\beta}_A, \widehat{\beta}_P, \widehat{\beta}_G)$ deviate from the nominal 95% level by one or two percentage points in Figure 1.4. It is reassuring to notice that this is simply due to the Monte Carlo error of the 100 simulations, because all the coverage rates converge to 95% as we increase the number of simulations to 1000, now that the assumptions of the benchmark estimators are satisfied here.

1.5.3 Results of diagnostic test

The results of the diagnostic test for the NILE assumption are given in Figure 1.5. Under each scenario, the histogram of the test statistic values over 100 simulations are compared to the χ^2 density function with 2 degrees of freedom, which is the distribution under the null hypothesis. At the 5% significance level, the rejection rate over the 100 simulations is 0.63 under Scenario-I, 0.06 under Scenario-II and 0.02 under Scenario-III.

Take first Scenario-III, where the set-up satisfies both the NILE assumptions and the regression model, the histogram of the test statistic values agrees reasonably well with its

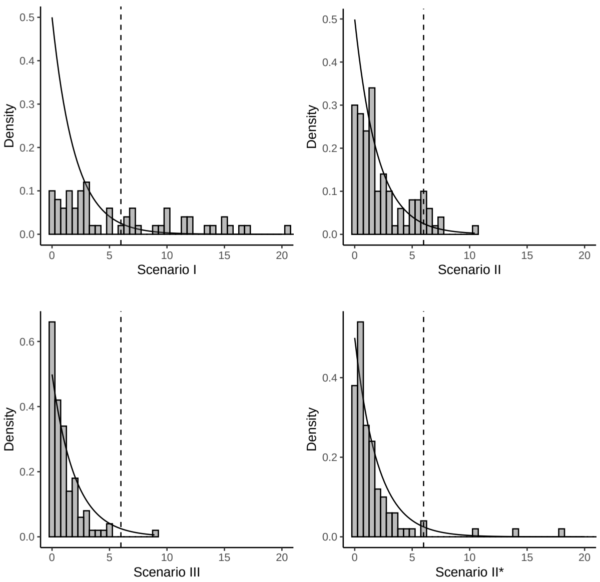


Figure 1.5. Diagnostic test for NILE assumption under Scenario-I to III: χ^2_2 density function (solid) with 95^{th} percentile (vertical dashed), histogram of observed test values over 100 simulations. Additional Scenario-II* with rejection ratio 0.04 over 100 simulations.

theoretical null distribution. Provided the relevant NILE assumptions, the higher missing-match rate of Gold linkage and the heterogenous linkage errors of sub-Gold linkage on expectation do not lead to unbalanced selection of the linked entities under either. Over the 100 simulations, the rejection rate of the diagnostic test at the 5%-level is 0.02, which appears to agree with the actual performance of $\widehat{\beta}_P$ and $\widehat{\beta}_G$.

Next, the set-up of Scenario-II satisfies the regression model assumptions, but it does not necessarily fulfil the NILE assumption *a priori*, since the errors of the key variables had been generated in ways which imitate real-life idiosyncrasies that are beyond our control. However, over the 100 simulations, the empirical SE of $\widehat{\beta}_G - \widehat{\beta}_P$ are 0.069 and 0.130 for the difference of intercept and slope, respectively, whereas the average of the corresponding SE estimates are 0.066 and 0.116. The histogram of the test statistic values agrees fairly well with its theoretical null distribution. The rejection rate of the 5%-level test is 0.06, which again seems reasonable in light of the actual performance of $\widehat{\beta}_P$ and $\widehat{\beta}_G$ in Figure 1.4. These are evidences suggesting that the relevant NILE assumptions can be met at least approximately in many practical situations.

Meanwhile, the test performance deteriorates under Scenario-I with real-life data for regression. For instance, the histogram of the test statistic values does not agree at all with the theoretical null distribution. The rejection rate of the 5%-level test is 0.63, which is unnecessarily high in light of the bias reduction that can be achieved by $\widehat{\beta}_P$ and $\widehat{\beta}_G$ here. The imbalance of regression outliers between the two linkage sets causes severe under-estimation of $V(\widehat{\beta}_G - \widehat{\beta}_P)$. For example, the empirical SE is 2452.5 for the difference in intercept estimates and 0.113 for the slope difference, but the average of the corresponding SE estimates is only 446.7 and 0.015, respectively. The severely under-estimated denominator of the test statistic (1.9) leads then to the high rejection rate over the simulations.

For confirmation we carried out additional simulations, where we simulated the 3-block ELE linkage errors, while retaining the real-life income data for regression. The results are shown in Figure 1.5, designated as Scenario-II*, which are similar to those under Scenario-II and III. The empirical SE is 1380.8 for the intercept difference and 0.061 for the slope, while the average of SE estimates is 1721.1 and 0.072, respectively, despite the presence of regression outliers. The rejection rate of the 5%-level test is now 0.04, which would be more helpful in practice. The cause of these results lies in the different set-ups of Scenario-I and II*. Although regression outliers are present in both cases, the linkage errors are randomly 'assigned' to the sample units under Scenario-II*, such that they may affect 'evenly' $\widehat{\beta}_P$ and $\widehat{\beta}_G$ over repeated simulations. Under Scenario-I, however, the linkage key variables are fixed for each individual, such that e.g. regression outliers affect only $\widehat{\beta}_P$ but not $\widehat{\beta}_G$, provided the outliers in the population happen to be associated only with sub-Gold linkage individuals but none of the Gold linkage individuals. Such peculiarities in the *fixed* population of regression and key variables can affect the simulation results unexpectedly.

Finally, taking altogether the test results from the simulation study here, it would seem reasonable that one should not interpret the p-value of the diagnostic test too stringently in practice, e.g. despite the p-value is only 0.05 or even slightly lower in a given situation, the estimators assuming non-informative linkage errors are still likely be very helpful.

1.6 Concluding remarks

Heterogenous linkage errors and incomplete match space are likely to prevail in most applications of record linkage. We propose a practical approach to linkage-data regression for secondary analysis, which can accommodate both in a simple manner, provided suitable NILE assumptions of the linkage errors. Application and simulation suggest that the relevant assumptions can be met at least approximately in many situations. In the simulation studies where the match space is incomplete, the proposed Pseudo-OLS method is more efficient than the existing adjustment methods that operate under the approximate assumption of complete match space. Moreover, we construct for the first time an accompanying diagnostic test for the NILE assumptions, which can provide helpful guidance in practice. Regarding future development, we believe additional research is needed for robust variance estimation, which can better cope with heterogeneous regression errors and potential outliers. As another current research topic we are developing an extension of our approach to categorical linkage-data analysis.

Appendix

Proof of Proposition 2

Provided (p1), i.e. (1.4) over D_1, we have $\sum_{i \in D_1^*} x_i x_i^\top / N^* - \sum_{i \in D_1} x_i x_i^\top / N_1 \xrightarrow{P} 0$, by the same argument as in Section 1.2.3, where $z_i = x_i x_i^\top$ for $i \in D_1$. Provided (p0.2) in addition, we have $\sum_{i \in D_1^*} x_i x_i^\top / N^* - \sum_{i \in D_M} x_i x_i^\top / N_M \xrightarrow{P} 0$. Likewise, $\sum_{i \in D_1^*} x_i / N^* - \sum_{i \in D_M} x_i / N_M \xrightarrow{P} 0$, and $\sum_{i \in D_1^*} y_i^* / N^* - \sum_{i \in D_M} y_i / N_M \xrightarrow{P} 0$ by (p1), i.e. (1.4) over D_2, and (p0.3). Notice that the conditions (p0.2) and (p0.3) are needed to ensure that false links of the unmatched records do not cause asymptotic bias to the 'marginal' statistics, i.e. $\sum_{i \in D_1^*} x_i x_i^\top / N^*$ and $\sum_{i \in D_1^*} x_i / N^*$ based on D_1 and $\sum_{i \in D_1^*} y_i^* / N^*$ based on D_2. Finally, provided (p1), i.e. (1.2) over D_1^*, and (p0.1), we have, as discussed in Section 1.2.4,

$$\begin{aligned}\sum_{i\in D_1^*} Cov(x_i, y_i^*|\ell_i = 1) &= \sum_{i\in D_{MM}^*} Cov(x_i, y_i) + \sum_{i\in D_{1M}^*\setminus D_{MM}^*} 0 + \sum_{i\in D_1^*\setminus D_{1M}^*} 0 \\ &= \sum_{i\in D_M} (\ell_i a_{ii}) Cov(x_i, y_i).\end{aligned}$$

Let $z_i = Cov(x_i, y_i)$ for $i \in D_M$, which is an unknown constant associated with $i \in D_M$. Now that the inclusion probability of $i \in D_{MM}^*$ from D_M is $\lambda_i\psi_i$, (p1) entails asymptotic NILE for $\ell_i a_{ii}$ over D_M, such that $\sum_{i\in D_{MM}^*} z_i/N_{MM}^* - \sum_{i\in D_M} z_i/N_M \xrightarrow{P} 0$. Notice that each term of S_{xy^*} from D_{MM}^* is an asymptotically unbiased estimate of the corresponding $Cov(x_i, y_i)$, and each term outside of D_{MM}^* has asymptotic expectation zero, so that $\lambda^{-1}S_{xy^*} - S_{xy}(M) \xrightarrow{P} 0$ given (p2), where $S_{xy}(M)$ is the empirical covariance of (x_i, y_i) over D_M. The consistency of $\widehat{\lambda}$ implies then $\widehat{\beta}_P - \tilde{\beta} \xrightarrow{P} 0$.

Approximate variance $V(\bar{x}\bar{y}^* + \widehat{\lambda}^{-1}S_{xy^*})$

We have $V(\bar{x}\bar{y}^*|A) = \bar{x}\bar{x}^\top\sigma^2/N^*$ and $E(\bar{x}\bar{y}^*|A) = \bar{x}\bar{x}^{*\top}\beta$, where $\bar{x}^* = \sum_{j\in D_2^*} x_j/N^* \neq \bar{x} = \sum_{i\in D_1^*} x_i/N^*$. Conditional on all the x's, we obtain

$$V(\bar{x}\bar{y}^*) = E[V(\bar{x}\bar{y}^*|A)] + V[E(\bar{x}\bar{y}^*|A)] = \bar{x}\bar{x}^\top\sigma^2/N^*.$$

Next, let $\nu_1 = V\big(\sum_{i\in D_1^*}(x_i - \bar{x})(y_i^* - \bar{y}^*)\big)$, where

$$\begin{aligned}\nu_1 &= \sum_{i\in D_1^*} (x_i - \bar{x})(x_i - \bar{x})^\top V(y_i^* - \bar{y}^*) + \sum_{i\in D_1^*}\sum_{k\neq i}(x_i - \bar{x})(x_k - \bar{x})^\top Cov(y_i^* - \bar{y}^*, y_k^* - \bar{y}^*) \\ &= \sum_{i\in D_1^*} (x_i - \bar{x})(x_i - \bar{x})^\top(1 - \frac{1}{n})\sigma^2 - \sum_{i\in D_1^*}\sum_{k\neq i}(x_i - \bar{x})(x_k - \bar{x})^\top\frac{1}{n}\sigma^2 \\ &= \sum_{i\in D_1^*} (x_i - \bar{x})(x_i - \bar{x})^\top\sigma^2 - \frac{\sigma^2}{n}\sum_{i\in D_1^*}(x_i - \bar{x})\sum_{k\in D_1^*}(x_k - \bar{x})^\top \\ &= N^*S_{xx}\sigma^2, \qquad \text{for} \quad S_{xx} = \frac{1}{N^*}\sum_{i\in D_1^*}(x_i - \bar{x})(x_i - \bar{x})^\top,\end{aligned}$$

since $\sum_{k\in D_1^*}(x_k - \bar{x}) = 0$, such that $V(\widehat{\lambda}^{-1}S_{xy^*}|A) = S_{xx}\sigma^2/(\widehat{\lambda}^2N^*)$. Moreover,

$$E(\widehat{\lambda}^{-1}S_{xy^*}|A) = \widehat{\lambda}^{-1}\frac{1}{N^*}\sum_{i\in D_1^*}(x_i - \bar{x})E(y_i^* - \bar{y}^*|A) = \widehat{\lambda}^{-1}S_{xx^*}\beta \approx \widehat{\lambda}^{-1}\lambda S_{xx}\beta,$$

where $S_{xx^*} = \sum_{i\in D_1^*}(x_i - \bar{x})(x_{j_i} - \bar{x}^*)^\top/N^*$, and x_{j_i} is the x-vector for y_j that is linked to the record i in D_1, which is uncorrelated to x_i unless $j_i = i$. Therefore, asymptotically as $|M| \to \infty$, we have $S_{xx^*} \approx S_{xx}N_{MM}^*/N^* \approx \lambda S_{xx}$. We obtain

$$V(\widehat{\lambda}^{-1}S_{xy^*}) = E\Big(\frac{\sigma^2}{\widehat{\lambda}^2N^*}S_{xx}\Big) + V\big(\widehat{\lambda}^{-1}\psi S_{xx}\beta\big) \approx \frac{\sigma^2}{\lambda^2N^*}S_{xx} + V(\widehat{\lambda})S_{xx}\beta\beta^\top S_{xx}^\top.$$

Putting together $V(\bar{x}\bar{y}^*)$ and $V(\widehat{\lambda}^{-1}S_{xy^*})$ from above, we have

$$\begin{aligned}
&V(\bar{x}\bar{y}^* + \widehat{\lambda}^{-1}S_{xy^*}) \approx (\bar{x}\bar{x}^\top + S_{xx})\frac{\sigma^2}{N^*} + \Delta = (\frac{1}{N^*}X_{D_1^*}^\top X_{D_1^*})\frac{\sigma^2}{N^*} + \Delta,\\
&\Delta = (\frac{1}{\lambda^2} - 1)\frac{\sigma^2}{N^*}S_{xx} + V(\widehat{\lambda})S_{xx}\beta\beta^\top S_{xx}^\top,\\
&V(\widehat{\beta}_P) = (X_{D_1^*}^\top X_{D_1^*})^{-1}\sigma^2 + (\frac{1}{N^*}X_{D_1^*}^\top X_{D_1^*})^{-1}\Delta(\frac{1}{N^*}X_{D_1^*}^\top X_{D_1^*})^{-1}.
\end{aligned}$$

Covariance of $\widehat{\beta}_G$ and $\widehat{\beta}_P$

The estimator $\widehat{\beta}_G$ given by (1.5) can be rewritten as

$$\begin{aligned}
&\widehat{\beta}_G = H_G(\bar{x}_G\bar{y}_G + \frac{1}{N_G^*}\tau_G) \qquad \bar{x}_G = \frac{1}{N_G^*}\sum_{i\in D_G^*} x_i \qquad \bar{y}_G = \frac{1}{N_G^*}\sum_{i\in D_G^*} y_i\\
&H_G = (\frac{1}{N_G^*}\sum_{i\in D_G^*} x_i x_i^\top)^{-1} \qquad \tau_G = \sum_{i\in D_G^*}(x_i - \bar{x}_G)(y_i - \bar{y}_G) = \sum_{i\in D_G^*}(x_i - \bar{x}_G)y_i
\end{aligned}$$

By definition we have $D_G^* \subset D_P^*$ and $N_G^* < N_P^*$. Let $D_A^* = D_P^* \setminus D_G^*$ consist of the remaining entities. Let $w = N_G^*/N_P^*$, and $1 - w = N_A^*/N_P^*$. We have

$$\begin{aligned}
&\widehat{\beta}_P = H_P(\bar{x}_P\bar{y}_P + \frac{1}{\widehat{\lambda}N_P^*}\tau_P) \qquad \bar{x}_P = \frac{1}{N_P^*}\sum_{i\in D_P^*} x_i \qquad H_P = (\frac{1}{N_P^*}\sum_{i\in D_P^*} x_i x_i^\top)^{-1}\\
&\bar{y}_P = \frac{1}{N_P^*}\sum_{i\in D_P^*} y_i^* = w\bar{y}_G + (1-w)\bar{y}_A^* \qquad \bar{y}_A^* = \frac{1}{N_A^*}\sum_{i\in D_A^*} y_i^*\\
&\tau_P = \sum_{i\in D_P^*}(x_i - \bar{x}_P)(y_i^* - \bar{y}_P^*) = \sum_{i\in D_P^*}(x_i - \bar{x}_P)y_i^* = \tau_G' + \tau_A\\
&\tau_G' = \sum_{i\in D_G^*}(x_i - \bar{x}_P)y_i \qquad \tau_A = \sum_{i\in D_A^*}(x_i - \bar{x}_P)y_i^*
\end{aligned}$$

Notice that $\tau_G \neq \tau_G'$ because τ_G involves $\bar{x}_G$ whereas τ_G' involves $\bar{x}_P$. Now, to obtain the covariance, we only need to take the cross terms one by one. We have

$$Cov(H_G\bar{x}_G\bar{y}_G, H_P\bar{x}_P\bar{y}_P) = wH_G\bar{x}_G V(\bar{y}_G)\bar{x}_P^\top H_P^\top = \frac{\sigma^2}{N_P^*}H_G\bar{x}_G\bar{x}_P^\top H_P$$

because $Cov(\bar{y}_G, \bar{y}_A^*) = 0$ and $H_P = H_P^\top$. Similarly, $Cov(\bar{y}_G, \tau_A) = 0$, such that

$$\begin{aligned}
&Cov(H_G\bar{x}_G\bar{y}_G, \frac{1}{\widehat{\lambda}N_P^*}H_P\tau_P) = E(\frac{1}{\widehat{\lambda}N_P^*})Cov(H_G\bar{x}_G\bar{y}_G, H_P\tau_G')\\
&\qquad \approx \frac{\sigma^2}{\lambda N_P^*}H_G\bar{x}_G(\bar{x}_G - \bar{x}_P)^\top H_P^\top = \frac{\sigma^2}{\lambda N_P^*}H_G\bar{x}_G\bar{x}_G^\top H_P - \frac{\sigma^2}{\lambda N_P^*}H_G\bar{x}_G\bar{x}_P^\top H_P\\
&Cov(H_P\bar{x}_P\bar{y}_P, \frac{1}{N_G^*}H_G\tau_G) = \frac{w\sigma^2}{N_G^*}H_P\bar{x}_P(\bar{x}_G - \bar{x}_G)^\top H_G^\top = 0
\end{aligned}$$

Finally, let $S_G^2 = \sum_{i \in D_G^*} (x_i - \bar{x}_G)(x_i - \bar{x}_G)^\top / N_G^*$, such that

$$Cov(\tau_G, \tau_G') = \sigma^2 \sum_{i \in D_G^*} (x_i - \bar{x}_G)(x_i - \bar{x}_P)^\top = \sigma^2 N_G^* S_G^2$$
$$Cov(\frac{1}{N_G^*} H_G \tau_G, \frac{1}{\widehat{\lambda} N_P^*} H_P \tau_P) = E(\frac{1}{\widehat{\lambda} N_P^* N_G^*}) H_G Cov(\tau_G, \tau_G') H_P^\top \approx \frac{\sigma^2}{\lambda N_P^*} H_G S_G^2 H_P$$

Summarising the four terms above, and noting $H_G = \bar{x}_G \bar{x}_G^\top + S_G^2$, we obtain

$$\begin{aligned} Cov(\widehat{\beta}_G, \widehat{\beta}_P) &\approx \frac{\sigma^2}{\lambda N_P^*} H_G (\bar{x}_G \bar{x}_G^\top + S_G^2) H_P + (1 - \frac{1}{\lambda}) \frac{\sigma^2}{N_P^*} H_G \bar{x}_G \bar{x}_P^\top H_P \\ &= \frac{\sigma^2}{\lambda N_P^*} H_P - (\frac{1}{\lambda} - 1) \frac{\sigma^2}{N_P^*} H_G \bar{x}_G \bar{x}_P^\top H_P \end{aligned}$$

Chapter 2

Bayesian analysis of one–inflated models for elusive population size estimation

Abstract

The identification and treatment of “one–inflation” in estimating the size of an elusive population has received increasing attention in capture–recapture literature in recent years. The phenomenon occurs when the number of units captured exactly once clearly exceeds the expectation under a baseline count distribution. Ignoring one–inflation has serious consequences for estimation of the population size, which can be drastically overestimated. In this Chapter we propose a Bayesian approach for Poisson, Geometric and Negative Binomial one–inflated count distributions. Posterior inference for population size will be obtained applying a Gibbs sampler approach. We also provide a Bayesian approach to model selection. We illustrate the proposed methodology with simulated and real data and propose a new application in official statistics to estimate the number of people implicated in the exploitation of prostitution in Italy.

2.1 Introduction

A popular methodology to estimate the size of an elusive population is the capture-recapture method, originally used to estimate animal abundance. When the captures are continuously collected over a fixed interval of time, and time is considered uninfluential, the total number of captures for each unit is the sufficient statistic. Here we focus on this setting, usually called “repeated counting data” Böhning & Schön (2005). To estimate the population size, the observation/capturing counting process must first be modelled.

In Farcomeni & Scacciatelli (2013), “one–inflation” is explicitly mentioned for criminal populations as a (simple) particular case in a broader class of behavioural effects. In more recent years, a series of papers (see, e.g., Godwin & Böhning (2017), Godwin (2017), Godwin (2019), Böhning et al. (2018), Böhning & Friedl (2021)), has been devoted specifically to the phenomenon in repeated counting data.

One–inflation consists in an excess of “ones” in the observed data, i.e., more units than expected are captured exactly once. The excess of “ones” is usually evaluated with respect to a chosen family of counting distributions: Godwin & Böhning (2017) considered one–inflation with respect to a “base” Poisson model, while Böhning & Friedl (2021) analyzed the inflation in the Geometric case. One–inflated Negative Binomial was introduced in Godwin (2017), and the finite mixture of one–inflated Poissons in Godwin (2019).

One–inflation can occur for different reasons; for instance, when some units of the population can no longer be captured after the first capture. Such may be the case of some wild animal populations. In fact, animals experiencing a capture may find it so unpleasant that some develop the will and ability to avoid subsequent captures. Much the same mechanism may also occur in human populations, particularly when the first capture is a matter of law enforcement, involves imprisonment or reveals an undesirable characteristic/behaviour. See Godwin & Böhning (2017) for ample discussion of the justifications and conditions for one–inflation in capture–recapture, also including an interpretation of one–inflation as limiting case of the so–called “trap shy” behavioural model (see, e.g., pg. 37 of McCrea & Morgan (2014) or pg. 119 of Borchers et al. (2002)). One–inflation deserves specific attention due to its effect on population size estimators. In fact, when not taken into account, one–inflation causes overestimation of the total population size. This also applies to the well–known lower–bound Chao estimator, as discussed in Chiu & Chao (2016) and Böhning et al. (2018).

In this Chapter we propose a Bayesian approach for counting data models with one–inflation. The properties of our models are analyzed with both simulation studies and real data applications. In particular, we apply our models to real data to estimate the size of some illegal populations active in Italy in 2014 and some real data available from the literature on capture-recapture, where the issue of one–inflation has been recognised.

The Chapter is organized as follows: in Section 2.2 we introduce the notation for repeated counting data and broadly illustrate Bayesian inference for population size with this kind of data. We describe the general model for one–inflated count data under an unspecified counting distribution and outline a Gibbs sampler algorithm to handle the one–inflated models. We also introduce a formal Bayesian procedure for model comparison in the presence of one–inflated models. Section 2.3 specifies the results under the Poisson and Geometric assumptions, corroborating our proposal with a simulation study. In Section 2.4 we introduce the Negative Binomial distribution and its one–inflated counterpart discussing the boundary problem via a simulation study. In Section 2.5 we illustrate some applications to real cases: first we show the results of our inference on data on prostitution exploitation in Italy in 2014; moreover, we apply our models to some popular datasets in capture–recapture literature. Section 2.6 concludes the Chapter with some remarks and discussion of open issues for further investigation.

2.2 Bayesian inference for population size

According to the standard formulation, consider a closed population (no births, deaths or migration) of size N. For each unit in the population, let Y be a random variable taking value $j = 0, 1, 2, \dots$ if the individual is observed/captured j times. We only observe the n individuals, $n \leq N$, which are captured at least once. Let $\mathbf{y} = (y_1, \dots, y_n)$ be the vector of the individual number of captures. Note that $\mathbf{y}$ will denote the result of the capture-recapture experiment which comprises both the number n of captured individuals and the number of captures for each observed individual.

Let n_j denote the number of individuals observed j times, that is, n_j is the frequency of count j in sample $\mathbf{y}$. Our interest is to estimate the number of uncaptured units n_0, and, consequently, the total population size $N = n + n_0$, on the basis of some model for the observed n_j.

Bayesian inference for the population size N can be obtained with standard Markov Chain Monte Carlo (MCMC) algorithms. In fact, let $f(y|\theta) = P(Y = y|\theta)$ for $y = 0, 1 \dots,$ be the probability distribution function for Y. The generic expression for the likelihood $f(\mathbf{y}|\theta, N)$ is

$$f(\mathbf{y}|\theta, N) = \binom{N}{n} f(0|\theta)^{N-n} \prod_{i=1}^{n} f(y_i|\theta). \tag{2.1}$$

Assuming independent priors for θ and N, i.e., $p(\theta, N) = p(\theta)p(N)$, the posterior distribution $p(\theta, N|\mathbf{y})$ can easily be drawn by, for example, updating the conditional distributions

$$p(\theta|N, \mathbf{y}) \propto f(0|\theta)^{N-n} \prod_{i=1}^{n} f(y_i|\theta)\, p(\theta)$$

and

$$p(N|\theta, \mathbf{y}) \propto \binom{N}{n} f(0|\theta)^{N-n}\, p(N).$$

We can generate from those posteriors via Gibbs or Metropolis-Hastings steps, according to the parametric family for Y and the prior for N.

In the Bayesian literature, common choices for the (default or non–informative) prior over N are:

- $p(N) \propto N^l$ for $l \in \{-2, -1, -1/2, 0\}$ possibly truncating the prior to an opportune upper bound; $l = -1$ corresponds to the Jeffreys' prior which is improper;

- Rissanen's prior (Rissanen (1983)) which is always proper and is given by $p(N) \propto 2^{-\log^*(N)}$, where $\log^*(N)$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2(\log_2(N)), \ldots\}$.

See Tardella (2002), Wang et al. (2007) and Xu et al. (2014) for extensive simulation studies. Note that

i) by assuming $p(N) \propto 1/N$, the full conditional distribution of $n_0 = N - n$ is Negative Binomial with size parameter n and probability $f(0|\theta)$ whatever the model for Y may be;

ii) the full conditional of θ corresponds to its posterior distribution when the zero counts are also known.

For example, when Y is Poisson(λ) and a priori we take the conjugate prior for λ which is Gamma($\alpha_\lambda, \beta_\lambda$) the latter step consists solely in the generation of a Gamma distribution with parameters given by $\alpha_\lambda + s$ and $\beta_\lambda + n + n_0$, where s is the sum of the observed captures. Similarly, when Y is Geometric(p) and a priori we take the conjugate prior for p which is Beta(α_p, β_p) this step consists in the generation of a Beta distribution with parameters $\alpha_p + n + n_0$ and $\beta_p + s$.

2.2.1 One–inflated models

We assume that in our population a specific behavioural mechanism is at work, by virtue of which an individual that would otherwise face multiple captures now has a positive probability ω of being captured just once.

Let Y denote the observed number of captures for a unit, and Y^* the latent value we would observe without the behavioural mechanism. The two variables are linked by means of the following infinite transition matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & \omega & 1-\omega & 0 & 0 & \cdots \\ 0 & \omega & 0 & 1-\omega & 0 & \cdots \\ 0 & \omega & 0 & 0 & \ddots & \\ \vdots & \vdots & \vdots & \vdots & & \end{pmatrix},$$

where the (k, j)–th element represents the conditional probability $P(Y = j - 1 \mid \omega, Y^* = k - 1)$. When $k > 1$ these conditional probabilities can be written as

$$P(Y = j \mid \omega, Y^* = k) = \omega^{(1-\delta_k(j))}(1-\omega)^{\delta_k(j)} \quad j = 1, k.$$

where $\delta_k(j)$ is Kronecker delta.

Let $f(k|\theta) = P(Y^* = k \mid \theta)$ be the probability distribution, depending on a given parameter, θ, of the number of captures without the behavioural effect, and let $F(\theta)$ denote the associated c.d.f. Then, the resulting distribution for Y is the one–inflated model defined as follows:

$$P(Y = j \mid \theta, \omega) = \begin{cases} f(0|\theta) & \text{if } j = 0; \\ (1-\omega)f(1|\theta) + \omega(1 - f(0|\theta)) & \text{if } j = 1; \\ (1-\omega)f(j|\theta) & \text{if } j > 1. \end{cases}$$

The conditional distribution of Y^* when $Y = j$ is concentrated on j when $j \neq 1$, while, when $j = 1$, we have:

$$P(Y^* = k \mid Y = 1, \theta, \omega) = \begin{cases} 0 & \text{if } k = 0; \\ \dfrac{f(1|\theta)}{f(1|\theta) + \omega(1 - F(1|\theta))} & \text{if } k = 1; \\ \dfrac{\omega f(k|\theta)}{f(1|\theta) + \omega(1 - F(1|\theta))} & \text{if } k > 1. \end{cases} \tag{2.2}$$

2.2.2 Gibbs sampler for one–inflated models

Bayesian inference for one–inflated models can be obtained by simulating the posterior distribution of $\theta, \omega, N, y_1^*, \ldots, y_n^*$ given the observed data $\mathbf{y}$, where $y_1^*, \ldots, y_n^*$ indicate the unknown captures that the n observed units would have faced without the behavioural mechanism. Let us assume that the parameters θ, ω and N are a priori independent and let $p(\theta, \omega, N) = p(\omega)p(\theta)p(N)$ denote the prior distribution. The general expression for the posterior distribution of one–inflated models augmented with the vector $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ is

$$\begin{aligned} p(\theta, \omega, N, \mathbf{y}^*|\mathbf{y}) &\propto p(\mathbf{y}|\theta, \omega, N, \mathbf{y}^*)p(\mathbf{y}^*, \theta, \omega, N) \\ &\propto \prod_{i=1}^{n} P(Y_i = y_i|y_i^*, \omega)p(\mathbf{y}^*|N, \theta)p(\theta)p(\omega)p(N) \\ &\propto \binom{N}{n} f(0|\theta)^{N-n} \prod_{i=1}^{n} P(Y_i = y_i|y_i^*, \omega)f(y_i^*|\theta)p(\theta)p(\omega)p(N). \end{aligned}$$

To describe our approach to simulate the posterior distribution of one–inflated models, we introduce an additional latent binary variable Z_i indicating the presence/absence of the behavioural mechanism which causes the one–inflation in unit i, i.e., Z_i is the indicator function of the event $\{Y_i \neq Y_i^*\}$. We then have that:

$$P(Z_i = 1 \mid Y_i \neq 1) = 0,$$

and, from (3.5), we have

$$P(Z_i = 1 \mid Y_i = 1) = \frac{\omega(1 - F(1|\theta))}{f(1|\theta) + \omega(1 - F(1|\theta))}.$$

Then, since $Z_i = 1$ implies $Y_i^* > 1$, we have

$$P(Y_i^* = k \mid Z_i = 1) = \begin{cases} \dfrac{f(k \mid \theta)}{1 - F(1 \mid \theta)} & \text{if } k > 1; \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

We can now outline a Gibbs sampler looping over the full conditionals of Y^* and ω, N and θ. The updating of θ will depend on the model assumption for Y^* and may require a Metropolis–within–Gibbs step, whereas the updating of Y^*, ω and N can always be performed with the following exact Gibbs steps:

i) The simulation of the full conditional of $Y_1^*, \ldots, Y_n^*$ can be obtained in two steps, by first updating $Z_1, \ldots, Z_n$. In fact, let $n_z = \sum_{i=1}^{n} Z_i$ be the number of units affected by one–inflation; then, conditional on the current value of ω and θ, we can generate a value for n_z from

$$Binom\left(n_1\,,\ \frac{\omega(1-F(1|\theta))}{f(1|\theta)+\omega(1-F(1|\theta))}\right).$$

Then, for each of the n_z units, we can generate a value of Y^* by simply simulating a number of captures from the truncated count distribution (3.7).

ii) Consider the prior

$$\omega \sim Beta(\alpha_\omega, \beta_\omega),$$

and let $n_{z,k}$ be the number of units among the n_z for which $Y^* = k$, such that $\sum_k n_{z,k} = n_z$. We can then write the full conditional of ω, $p(\omega \mid -)$ as:

$$p(\omega \mid -) \propto \omega^{\alpha_\omega - 1}(1-\omega)^{\beta_\omega - 1} \prod_{k>1} \left[\omega f(k \mid \theta)\right]^{n_{z,k}} \cdot \left[(1-\omega) f(k \mid \theta)\right]^{n_k}.$$

That is, we can directly draw ω from

$$Beta\left(\alpha_\omega + n_z\,,\ \beta_\omega + \sum_{k>1} n_k\right).$$

iii) The full conditional distribution of N is given by

$$p(N \mid -) \propto \binom{N}{n} f(0|\theta)^{N-n} p(N)$$

and, by assuming the improper prior $p(N) \propto 1/N$ we can directly draw n_0 from the following Negative Binomial

$$\binom{N-1}{n-1} f(0|\theta)^{N-n}(1-f(0|\theta))^n.$$

If we adopt a different prior over N, we have to implement a Metropolis step.

Finally, as we have seen, the updating of θ will depend on the model assumption for Y^*. The general expression for the full conditional of θ is:

$$p(\theta \mid -) \propto f(0|\theta)^{N-n} \prod_{i=1}^{n} f(Y_i^*|\theta) p(\theta).$$

2.2.3 Model selection

To test the one–inflation assumption with respect to a specific base count distribution we can adopt a fully Bayesian approach. Let M_1 be the non–inflated model and M_2 the one–inflated counterpart, (indicated by the OI suffix, hereafter). Model comparison can be performed by calculating the posterior model probabilities

$$P(M_i \mid \mathbf{y}) = \frac{p(M_i)p(\mathbf{y}|M_i)}{p(M_1)p(\mathbf{y}|M_1) + p(M_2)p(\mathbf{y}|M_2)}$$

where $p(\mathbf{y}|M_i)$ is the marginal likelihood that, for the models considered in this Chapter, can be generally written as

$$p(\mathbf{y}|M_i) = \int \sum_{N=n}^{\infty} f(\mathbf{y} \mid \theta_i, N, M_i) p(\theta_i, N \mid M_i)\, d\theta_i,$$

with θ_1 and θ_2 denoting respectively the parameters of the baseline and the OI counterpart models. For instance, for Poisson model we have $\theta_1 = \lambda$ and $\theta_2 = (\lambda, \omega)$, for the Geometric case we have $\theta_1 = p$ and $\theta_2 = (p, \omega)$. In the case of two models we can directly use the Bayes factor (BF) in favour of the OI

$$BF = \frac{P(M_2 \mid \mathbf{y})}{P(M_1 \mid \mathbf{y})} = \frac{P(M_2)}{P(M_1)} \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)}.$$

Note that we can also extend the comparison setting by simultaneously considering more than two models. For example, in the next Section we compare the Poisson and the Geometric model together with the corresponding OI counterparts for a total of 4 models. Assuming equal prior probabilities $P(M_i)$ for $i = 1, \ldots, k$, the posterior model probabilities are proportional to the marginal likelihoods, that is $P(M_i \mid \mathbf{y}) \propto p(\mathbf{y}|M_i)$ for $i = 1, \ldots, k$. Note. moreover, that assuming the non–informative prior $p(N) = c/N$ would produce marginal likelihoods depending on the constant c. However, in our case, the parameter N has the same meaning across all the models under comparison, hence the use of the same improper prior $p(N) = c/N$ is justified and the constant c cancels out in the evaluation of the posterior model probabilities, (see Kass & Raftery (1995)).

Analytical evaluation of the marginal likelihoods $p(\mathbf{y}|M_i)$ is not possible. However, we have that (see Appendix)

$$p(\mathbf{y}|M_i) = c \int \sum_{N=n}^{\infty} f(\mathbf{y}|\theta_i, N, M_i) \frac{1}{N} p(\theta_i)\, d\theta_i = \frac{c}{n} \int \prod_{i=1}^{n} \frac{f(y_i|\theta_i)}{1 - f(0|\theta_i)} p(\theta_i) d\theta_i. \quad (2.4)$$

Hence, the posterior model probabilities will depend solely on fitting the truncated distribution of Y to the observed captures.

To evaluate the marginal likelihood of each model numerically, we use the Chib's approximation introduced in Chib (1995) which can easily be obtained as a by-product of the general Gibbs algorithm illustrated in the previous Section. The details of the Chib approximation for all the models considered throughout this Chapter are given in the Appendix.

Finally, it is worth noting that, in the context of capture–recapture, model averaging appears to be a suitable alternative to model selection. In fact, the quantity of interest N has the same meaning across different models and we can easily obtain an estimate $\overline{N}$ of N averaged over the eligible alternatives via the following formula:

$$\overline{N} = E[N \mid \mathbf{y}] = \sum_i \widehat{N}_{M_i}\, P(M_i \mid \mathbf{y}),$$

where $\widehat{N}_{M_i}$ is the posterior mean of N obtained under model M_i. However, since the estimates of N under the base model and under its one–inflated counterpart may show very considerable differences, definite choice between the two could be a sensible approach in this case.

2.3 One–inflated Poisson and Geometric distributions

If we assume that our count data Y^* follows a Poisson distribution, i.e., $f(\theta)$ represents a Poisson density with parameter λ, the model proposed for the observed Y in previous section 2.2.1 is a one–inflated Poisson (OIP) and corresponds to the model presented in Godwin & Böhning (2017).

The estimating procedure is based on the Gibbs sampler described in Section 2.2.1, where, in order to complete the analysis framework, we assume a Gamma(α_λ,β_λ) prior for λ, α_λ and β_λ being shape and rate parameters. Let n_k^* be the total number of units captured k times after updating n_0, n_z and Y^*, that is,

$$n_k^* = \begin{cases} n_0 & \text{for } k = 0; \\ n_1 - n_z & \text{for } k = 1; \\ n_k + n_{z,k} & \text{if } k > 1; \end{cases}$$

and let $\{n^*\}$ denote the set of all values n_k^* for $k = 0, 1, ...$ We can then generate the updated value for λ from its full conditional

$$Gamma\left(\alpha_\lambda + \sum_{k>0} k\, n_k^*\ ,\ \beta_\lambda + N\right).$$

If we adopt a Geometric distribution for Y^*, parameterized as

$$P(Y^* = k \mid p) = (1-p)^k p,$$

the resulting model for Y is called one–inflated Geometric (OIG). To finalize the Bayesian analysis, we adopt a $Beta(\alpha_p, \beta_p)$ conjugate prior for p, and its posterior conditional on the current values of n_0, n_z and Y^* would be equal to:

$$Beta\left(\alpha_p + N\ ,\ \beta_p + \sum_{k>0} k\, n_k^*\right).$$

2.3.1 A simulation study

In this section we present a two–fold simulation study; on one hand, we aim to validate our proposal for inference on the population size in the presence of one-inflation, while on the other hand the results of the simulation study illustrate the model selection among the four models presented in the previous section, namely, Poisson (which we refer to as model Poi), Geometric (Geo), One–inflated Poisson (OIP), and One–inflated Geometric (OIG). Specifically, we set up three main scenarios: in the first we generate from the base distributions without one–inflation; in the second scenario, we generate from one–inflated distributions with a low/moderate inflation rate ($\omega = 0.2$), while in the third we consider a substantial inflation rate ($\omega = 0.5$). We repeat each scenario with 2 different values of the parameter (λ or p) and with 2 different values of N (500 and 1000). We set the parameters using values similar to those from the real cases analysed in Section 2.5. The scenarios and the values of the different parameters are summarised in Table 2.1.

For each combination of parameters in each scenario we simulate 100 datasets of N units from the respective generating model and remove the 0–counts from the sample. To simulate from the one–inflated models in Scenarios II and III, we generate from the corresponding base model and then change each generated value greater than 1 to a 1 with probability ω. All the experiments were conducted in R and the code is available as Supporting Information on the journal's web page.

Table 2.1. Simulation scenarios with data generating models, parameter values, and expected sample size $E[n]$ (The expected values of n are common to all three scenarios)

Scenario I No inflation	Scenario II Low inflation, $\omega = 0.2$	Scenario III Substantial inflation, $\omega = 0.5$	N	Distribution Parameter	$E[n]$
Poi	OIP	OIP	500	$\lambda = 1$	316
				$\lambda = 2$	432
			1000	$\lambda = 1$	632
				$\lambda = 2$	865
Geo	OIG	OIG	500	$p = 0.4$	300
				$p = 0.6$	200
			1000	$p = 0.4$	600
				$p = 0.6$	400

First, we set out to evaluate the sensitiveness of the estimates of the unobserved population size n_0 under mispecification of the model. For each simulated dataset, we consider the estimates of n_0, given by the posterior mean, under all four models, and compute relative bias calculated as the relative difference between the true value and the posterior mean of the parameter. As priors, we adopt quite non–informative choices: we set a uniform $\omega \sim Beta(1, 1)$ in all one–inflated models. We set $p \sim Beta(1, 1)$ in the Geometric and OIG models, and $\lambda \sim Gamma(0.01, 0.01)$ in the Poisson and OIP. Different values for the Gamma prior were also tested, obtaining very similar results. In fact, we use the same priors in all the scenarios, regardless the data generating models. Clearly this simulation setting does not exhaust the investigation of the priors' role on the results of the model selection, however, since we do not use the values from which we generated the data in any priors in any scenario, we believe that the simulation results are a good assessment of the proposed model selection criterion.

Table 2.2 shows the average percentage relative bias over the 100 replicates.

Table 2.2. Relative bias (%) of the unobserved units estimates, n_0

Generating Model			$N = 500$				$N = 1000$			
Model	Parameter	Inflation	Poi	Geo	OIP	OIG	Poi	Geo	OIP	OIG
Poi	1	None	1.67	198	-12	189	0.37	196	-9	190
Poi	2	None	1.28	391	-5.49	389	0.88	390	-4.12	388
Geo	0.4	None	-82	-0.80	-91	-5.48	-82	-1.13	-91	-4.33
Geo	0.6	None	-68	0.27	-80	-9.34	-68	0.73	-82	-6.84
OIP	1	0.2	52	514	3.41	501	52	514	2.32	507
OIP	2	0.2	37	273	0.71	246	37	272	0.38	254
OIP	1	0.5	147	497	14	339	146	496	6.04	146
OIP	2	0.5	218	883	5.38	619	219	886	3.54	614
OIG	0.4	0.2	-72	25	-91	0.92	-73	23	-91	-0.03
OIG	0.6	0.2	-55	26	-79	1.50	-56	26	-81	1.21
OIG	0.4	0.5	-39	100	-91	1.72	-39	100	-91	2.07
OIG	0.6	0.5	-16	108	-76	15	-18	104	-79	7.74

The results set out in Table 2.2 confirm that the estimates of n_0 we obtain with a one–inflated model are always lower than those obtained with the corresponding base model. In fact, ignoring one–inflation when present leads to severe and systematic overestimate of n_0.

On the other hand, admitting one–inflation when it is not present is not such a serious error and, on average, we moderately underestimate n_0. Choosing the wrong model (Poisson instead of Geometric, inflated or not) can have disastrous consequences. In particular, if data come from Poi or OIP models, a Geo or OIG models would drastically overestimate n_0. If data are generated from a Geo or OIG model, choosing a Poi or OIP model implies an equivalent underestimate of n_0. Note that, the two cases having the highest relative bias under the correct models can be justified by the observed number of captures. In particular, when the generating model is OIP with $\lambda = 1$ and $\omega = 0.5$, the expected number of captured units is low ($E[n] = 316$ when $N = 500$), and most of them are captured exactly once ($E[n_1] = 250$). The same happens in the case of OIG with $p = 0.6$ and $\omega = 0.5$ where $E[n] = 200$ and $E[n_1] = 160$. However, even in these worst cases, the relative bias decreases, as expected, when the sample size increases.

Here we will not present the results concerning the relative root mean squared error and the relative mean absolute error, which in any case, confirm the results presented on the relative bias.

These results are also confirmed on analysing the coverage of the posterior credible intervals, not reported here for brevity but computed by the R code available in the Supporting Information on the journal's web page. The posterior credible intervals of the one–inflated model almost always contain the true values when we generate from the corresponding baseline distribution. On the other hand, when we generate from a one–inflated model, the credible intervals of the baseline model barely cover the true values. The credible intervals deriving from the Poisson models (regardless of one–inflation) seldom cover the true value generated by the Geometric distribution, and vice-versa. The only exception is the case in which we generate from OIG ($p = 0.6$, $\omega = 0.5$) and estimate with a Poisson distribution (see the bottom row in Table 2.2), in which case the baseline Poisson credible intervals cover the true value nearly 50% of the times.

Next, to assess the model selection criterion detailed in the previous section, Figures 2.1 and 2.2 show the posterior probabilities of our four competing models calculated with Chib's approximation. Figure 2.1 summarizes the results in all the scenarios when $N = 500$, while Figure 2.2 refers to the case $N = 1000$.

It is evident that, as the number of observed units n increases, the effectiveness of the posterior model probabilities in identifying the correct generating model is reinforced. Note that n depends both on N and on the parameters λ and p. It is also evident that a higher inflation rate will be more easily identified correctly. In fact, when $N = 1000$, we would select the true data generating model in almost all simulations in Scenarios I and III, and in most cases in Scenario II. For the sake of brevity, here we do not present the results when $N = 2000$ or higher, since in all scenarios and parameter combinations the posterior model probability of the generating model is close to one.

When $N = 500$, we would still identify the correct generating model in the majority of cases, but we can observe some critical situations. In particular, when the generating model is OIP with $\lambda = 1$ and $\omega = 0.2$, and when we generate from the OIG with $p = 0.6$ and $\omega = 0.2$, the correct model and its base counterpart are almost equally preferable. In the former case we have $n = 316$ and $n_1 = 183$ on average, i.e., most of the units are captured once. Consequently, the posteriors probabilities are very similar due to such a slight alteration in singleton counts from the basic Poisson distribution. Much the same happens in the latter case, with an even lower number of observations (on average $n = 200$).

For a simulation study using frequentist criteria for model selection, (AIC and BIC) see Böhning & Ogden (2021).

In conclusion, as expected, the one–inflation models encompass the baseline models and, when one–inflation is not present, the slight underestimation of N decreases as n increases. Clearly, the choice of the distribution is a crucial aspect, and the Bayesian approach gives us a powerful tool to deal with model selection.

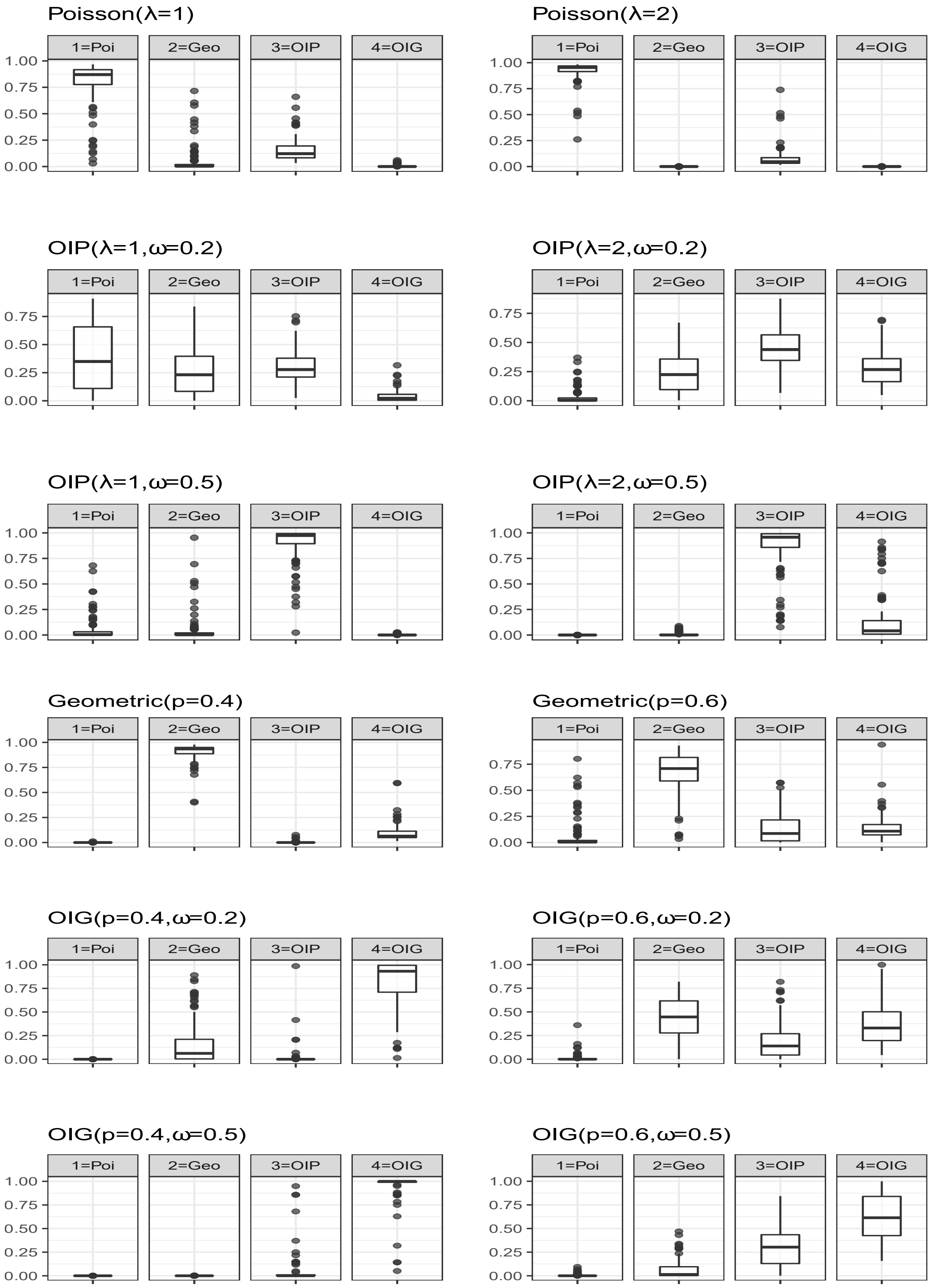


Figure 2.1. Box-plot of posterior model probabilities when $N = 500$; the data generating model is indicated above each panel

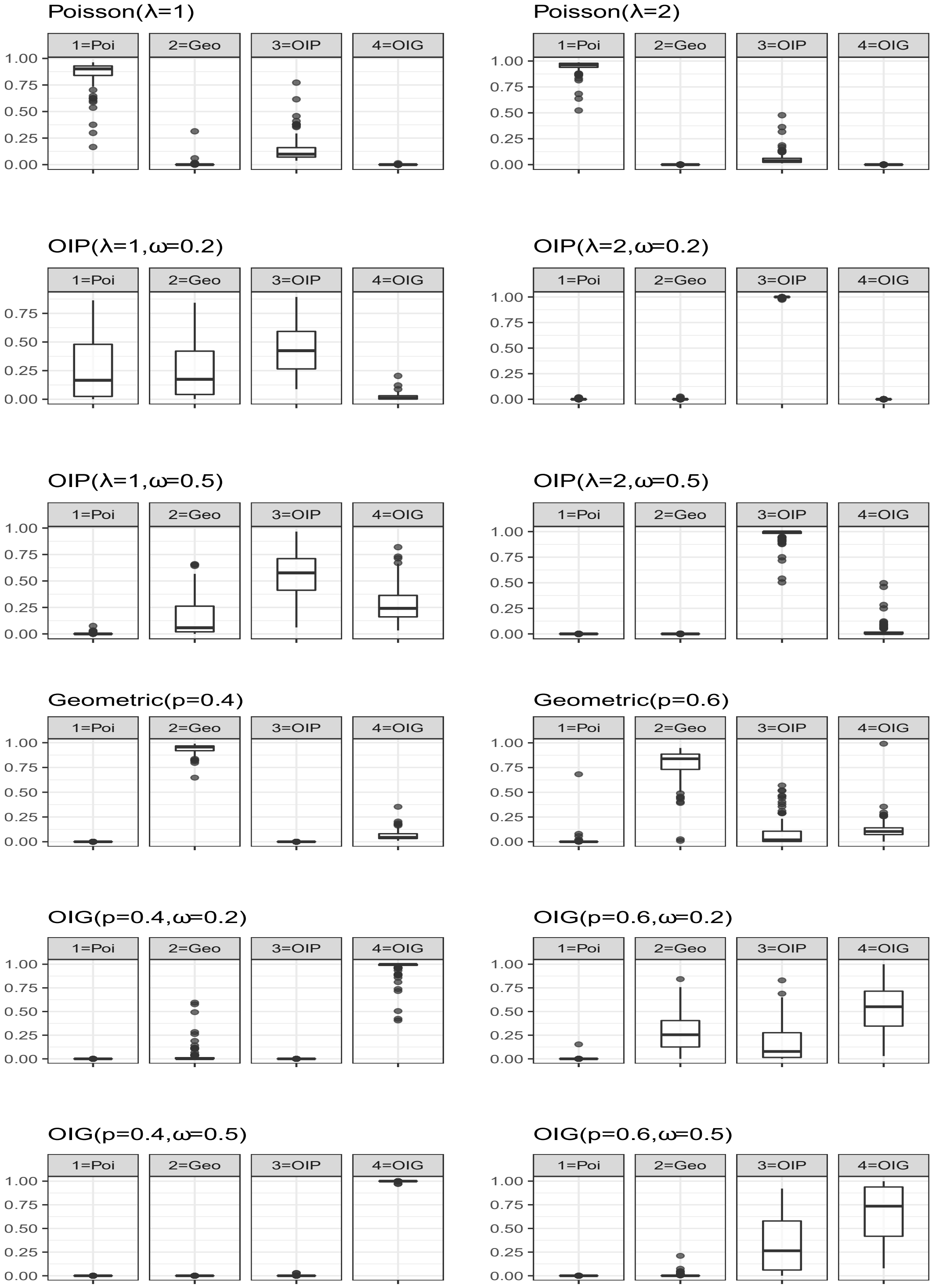


Figure 2.2. Box-plot of posterior model probabilities when $N = 1000$; the data generating model is indicated above each panel

2.4 One–inflated Negative Binomial

In this Section we describe how to perform Bayesian estimation of the population size in the presence of one-inflation when the base distribution is the Negative Binomial model. We also underline the inferential drawbacks related to this distribution which limit its general use and how the Bayesian approach mitigates these problems.

The Negative Binomial distribution (NB) is often adopted as a two-parameter generalization of Poisson that can take into account over-dispersed count data. It also constitutes a generalization of the Geometric distribution, with respect to which it allows for both overdispersion and underdispersion. Its use is well known in capture–recapture, and has also been investigated in the presence of one–inflation in Godwin (2017).

Here we assume that the unobserved count Y^* follows an NB model with the following parameterization in terms of r and p:

$$P(Y^* = k \,|\, r, p) = \frac{\Gamma(k+r)}{\Gamma(r)k!} p^r (1-p)^k, \tag{2.5}$$

and we will call the resulting model for Y One–inflated Negative Binomial (OINB). In our Bayesian approach, we set two independent priors on the parameters p and r. For p we take a $Beta(\alpha_p, \beta_p)$ prior, while for r we compare Gamma and Inverse Gamma priors in order to evaluate the different tail behaviour of these distributions on the posterior summaries.

The Gibbs sampler we developed follows the same passages presented in Section 2.2.1, where $f(\theta)$ takes the form (2.5). Recall that n_k^* represents the number of units captured k times after updating n_0, Z and Y^*. Then, generating from the full conditional of p presents no difficulties, as it turns out to be:

$$[p \,|\, -] \sim Beta\left(\alpha_p + Nr \;,\; \beta_p + \sum_{k>0} k\, n_k^*\right).$$

To update r, we compare two different approaches: a Gaussian random-walk Metropolis-Hastings step, and the two-stage Gibbs sampler proposed by Zhou & Carin (2015). Note also that the presence of a Metropolis step does not preclude calculation of the marginal likelihood $p(\mathbf{y}|M_i)$ with Chib's approximation for the Negative Binomial model and for the corresponding OI counterpart, as illustrated in Chib & Jeliazkov (2001). The Appendix provides details of the marginal likelihood approximation for these models.

2.4.1 Metropolis Hastings

The full conditional of r results in:

$$P(r \,|\, -) \propto p^{Nr} \prod_{k=0,1,\dots} \left(\frac{\Gamma(k+r)}{\Gamma(r)k!}\right)^{n_k^*} \frac{r^{\alpha_r - 1}}{e^{r\beta_r}}.$$

If we consider a Gaussian random walk Metropolis-Hastings, we accept a proposed value r' with probability equal to the minimum between 1 and

$$\exp\Big\{\sum_k n_k^* \left[\log\Gamma(r'+k) - \log\Gamma(r') - \log\Gamma(r+k) + \log\Gamma(r)\right] + N(r'-r)\log(p) + \Psi\Big\},$$

where

$$\Psi = \begin{cases} (\alpha_r - 1)\log(r'/r) + \beta_r(r - r') & \text{if } r \sim Gamma(\alpha_r, \beta_r); \\ (\alpha_r - 1)\log(r/r') + \beta_r(1/r - 1/r') & \text{if } r \sim InvGamma(\alpha_r, \beta_r). \end{cases}$$

2.4.2 Two-stage Gibbs sampler

Zhou & Carin (2015) exploit the representation of the Negative Binomial as a compound Poisson distribution, introduced by Quenouille (1949):

$$Y_i^* \sim \mathrm{NB}(r,p) \quad \Longleftrightarrow \quad Y_i^* = \sum_{j=1}^{l_i} u_{i,j}$$

where

$$l_i \sim Poisson(-r\log(p)) \qquad \text{and} \qquad u_{i,j} \overset{iid}{\sim} Logarithmic(1-p).$$

They found the explicit distribution of the full conditional of l_i to be the Chinese Restaurant Table (CRT) distribution with concentration parameter r. The two Gibbs steps are then:

i) We sample the latent counts, l_i, associated with each observed count y_i^*, which can be generated as:

$$l_i = \sum_{j=1}^{y_i^*} v_j, \qquad v_j \sim Bernoulli\left(\frac{r}{r+j-1}\right).$$

ii) We sample r from its full conditional which, given the conjugacy between the Gamma prior for r and the Poisson distribution, results in

$$[r \mid -] \sim Gamma\left(\alpha_r + \sum_{i=1}^{n} l_i \ , \ \beta_r - N\log(p)\right). \tag{2.6}$$

Note that, since the total number of captures is often in the order of thousands, and in (2.6) we are only interested in generating the sum of the l_i, we can simply adopt a Gaussian approximation in the first step. That is,

$$\sum_i l_i \sim N\left(\sum_i E[l_i], \sum_i Var[l_i]\right).$$

2.4.3 Boundary problem

The use of the NB in capture–recapture is limited by the so called "boundary problem" (see, e.g., Böhning (2015)). That is, when the estimate of r approaches zero, the Horvitz–Thompson estimation of the population size diverges. More generally, when in the observed (truncated) data the mean number of captures is close to one (which is typically the case in the presence of one–inflation), the NB model severely overestimates N, sometimes by several orders of magnitudes, even in simulated data generated by the NB itself. As pointed out in Godwin (2017), taking into account one–inflation alleviates this phenomenon, but does not completely avoid it.

We can confirm that, even in our Bayesian approach to the OINB model, we come up against the boundary problem. In general, we noted a great sensitivity of estimates of N to small differences in the value of parameter r, particularly when $r < 1$, and, accordingly, a great sensitivity of the estimates to specification of the prior distribution over r.

We see this phenomenon as an opportunity to investigate the usefulness of the Bayesian approach in further alleviating the boundary problem under the OINB. To this end, we conduct a simulation study to assess the effect of different prior specifications on the parameter r. We generate 100 replications of random values drawn from an OINB with

Table 2.3. Boundary cases for $\hat{r}$ and $\hat{N}$, %bias and %MSE of $\hat{N}$ for some prior specifications of r. Results from MLE in the bottom row, for comparison

N=5000				
Prior distribution of r	% Boundary cases for r	% Boundary cases for N	% bias of $\hat{N}$	% MSE of $\hat{N}$
Gamma(0.1,0.1)	33	30	218.59	1618.82
Gamma(1,1)	11	11	97.64	859.51
InvGamma(0.1,0.1)	0	0	-10.52	6.71
InvGamma(0.5,0.5)	0	0	-15.58	5.13
InvGamma(1,1)	0	0	-19.06	5.27
InvGamma(1,2)	0	0	-26.70	7.91
MLE	16	3	91.75	2217.32
N=500				
Prior distribution of r	% Boundary cases for r	% Boundary cases for N	% bias of $\hat{N}$	% MSE of $\hat{N}$
Gamma(0.1,0.1)	25	73	5043	1673356
Gamma(1,1)	0	8	249	7122
InvGamma(0.1,0.1)	0	0	-48	24
InvGamma(0.5,0.5)	0	0	-47	23
InvGamma(1,1)	0	0	-44	20
InvGamma(1,2)	0	0	-48	23
MLE	27	20	2422	584890

parameters $p = 0.35$, $r = 0.5$, and $\omega = 0.5$, and we go on to test two values for N, 5000 and 500. The observed sample size n varies at each replication; its expected value over the 100 replications is 2040, and 204 when $N = 5000$ and $N = 500$, respectively. The values of these parameters are comparable to the values studied in Godwin (2017), in the frequentist setting, and they allow us to mimic some real cases analysed in Section 2.5. All the experiments were conducted in R; the code is available as Supporting Information on the journal's web page.

We test some prior specifications on the r parameter, considering both the Gamma and the Inverse Gamma distributions. For estimation of r, we apply both the Metropolis-Hasting step and the two-stage Gibbs sampler proposed by Zhou & Carin (2015), observing negligible differences in the results. The outcomes presented in this Section are obtained using the Metropolis-Hasting approach. Finally, we compare the results with the maximum likelihood estimates for the OINB.

Table 2.3 shows the percentage relative bias and the percentage mean squared error (MSE) of the population size estimates, considering the difference between the true value and the mean of the posterior distribution obtained by the MCMC simulations. Table 2.3 also gives the number of cases, in percentage, where we encountered the boundary problem. In fact, we can define the boundary problem on both $\hat{r}$ and $\hat{N}$. We adopt the following convention: on $\hat{r}$, we set the boundary problem if $\hat{r} < 0.25$, while on $\hat{N}$, this is the case if $\hat{N} > 5N$. Finally, Table 2.3 presents the results of the maximum likelihood approach (MLE), obtained using the model proposed by Godwin (2017) and the R code provided by him as Supporting Information.

The Bayesian procedure implements the algorithm described in Section 2.4.1, setting the number of replications of the MCMC algorithm to $2 \cdot 10^6$. We set, a priori, $p(N) \propto 1/N$, and $Beta(1, 1)$ for both ω and p. From Table 2.3, it can be seen that a weakly informative prior specification for r, like $Gamma(1, 1)$ can already help reduce the boundary problem, when compared to the MLE approach. The boundary problem can be yet further limited using the Inverse Gamma as prior distribution for r. In the simulation, the Inverse Gamma prior has the double advantage of reducing both the boundary problem and the MSE of the estimates, at the cost of introducing a negative bias (underestimation) of the population size N, which is more severe for small Ns. Note that we used the convention of defining the occurrence of the boundary problem when $\hat{r} < 0.25$, while in Godwin (2017) the boundary problem is fixed at $\hat{r} < 0.05$. We believe that $\hat{r} < 0.25$ already suffices to indicate the presence of this phenomenon since, as clearly emerges from Table 2.3, it corresponds approximately to an estimate of N 5 times larger than its true value.

To further illustrate the performance of the NB and the OINB, with and without the boundary problem, we compare them with the models considered in Section 2.3 via a simulation study. In particular, we generate values from the NB with parameters $N = 5000$, $p = 0.35$, and from the OINB with parameters $N = 5000$, $p = 0.35$, and $\omega = 0.5$, under different scenarios for the size parameter r. For each scenario we generate 100 datasets and calculate the estimates of N given by the posterior mean, under the six models: Poisson, Geometric, Negative Binomial and their one–inflated counterparts. Table 2.4 shows the average percentage relative bias and relative mean squared error over the 100 replicates. As we have said, the value of the parameter r appears to be crucial in identifying the boundary problem for the NB model, and, under the OINB model, ω, too, has a clear role. As a consequence, the critical values for r differ under the two models. In our data generated from the NB, with the aforementioned values for p and N, we start to observe a substantial instability in the estimates when $r = 0.25$, and the sheer overestimation of N from the NB itself appears clearly in all simulations when $r = 0.1$ (not showed in the Table). When we generate from the OINB, estimates derived from the OINB itself start to show the same problem when $r = 0.5$.

We can see in Table 2.4 that, in the absence of the boundary problem, ($r = 1.5$ in both cases), the results confirm that the two models can be safely utilized if their respective model assumptions hold; in fact, they perform better than all other competing models. As already observed in Section 2.3, admitting one–inflation when it is not present leads to moderate underestimation, while ignoring one-inflation when present causes severe overestimation of N. In fact, in all cases, the NB overestimates N by several orders of magnitude with data generated from the OINB.

A counter-intuitive case is given by the data generated from the OINB with $r = 0.5$, in which case the OINB itself results as the second best model, the best being the non inflated Geometric. The explanation we gave to this result is the following: the Geometric model ignores one–inflation, and this fact should lead to an overestimation of N, but at the same time, it fixes the parameter r to 1, which is higher than the actual parameter of the generating model ($r = 0.5$), and this fact should imply an underestimation of N. Apparently, in our simulation, these two factors balance each other, giving the Geometric a better performance than the OIG and the OINB itself. In conclusion, when the model hypothesis are met, and the boundary problem is absent or not too serious, for values of r greater than 0.25 under the NB, and greater than 0.5 under the OINB, the use of an Inverse Gamma prior may alleviate the phenomenon. However, when the problem is evident, we advise against the use of the two models.

2.5 Results on estimating illegal populations

Illegal activities are by their very nature difficult to measure because the people involved have obvious reasons to hide them. In this Section, we apply our models to estimate the number of people implicated in the exploitation of prostitution, in Italy in 2014. In

Table 2.4. Results on %bias and %MSE of $\hat{N}$.

Generating model: OINB with $p = 0.35$ and $\omega = 0.5$

	$r = 0.5$ $(E[n] = 2040)$		$r = 1.5$ $(E[n] = 3695)$	
	% bias of $\hat{N}$	% MSE of $\hat{N}$	% bias of $\hat{N}$	% MSE of $\hat{N}$
Poi	-38.11	14.55	-7.25	0.54
Geo	5.19	0.38	42.31	17.94
NB (Gamma)	$4 \cdot 10^{13}$	$9 \cdot 10^{26}$	$4 \cdot 10^{11}$	$2 \cdot 10^{23}$
NB (Inv. Gamma)	2518	$2 \cdot 10^{5}$	$2 \cdot 10^{5}$	$2 \cdot 10^{10}$
OIP	-56.38	31.80	-19.32	3.74
OIG	-29.75	8.89	12.78	1.65
OINB (Gamma)	246	2898	1.81	0.25
OINB (Inv. Gamma)	-11.73	5.68	0.49	0.19

Generating model: NB with $p = 0.35$

	$r = 0.25$ $(E[n] = 1154)$		$r = 1.5$ $(E[n] = 3965)$	
	% bias of $\hat{N}$	% MSE of $\hat{N}$	% bias of $\hat{N}$	% MSE of $\hat{N}$
Poi	-71.04	50.48	-17.81	3.17
Geo	-53.99	29.17	10.98	1.21
NB (Gamma)	162.37	2044.18	0.19	0.03
NB (Inv. Gamma)	-9.64	5.06	0.02	0.03
OIP	-74.70	55.80	-19.16	3.67
OIG	-57.52	33.11	10.91	1.20
OINB (Gamma)	5.71	64.03	-1.58	0.05
OINB (Inv. Gamma)	-43.97	20.43	-1.86	0.06

addition, in Section 2.5.1 we illustrate the results obtained on some well-known data-sets in capture-recapture literature.

In Italy, prostitution is neither prosecuted nor regulated, but trafficking, exploitation, and aiding and abetting of prostitution is a crime subject to legal sanctions. These activities are mostly under the control of organised crime. In this study we exploit administrative records from the Ministry of Justice which report complaints, of victims or witnesses, for which the judicial authority has collected sufficient evidence to initiate a criminal proceeding.

On the basis of soft identifiers (date, country of birth and gender), the perpetrators can be identified and followed over a given time span, which is one year in this application. In this way, the administrative source can be viewed as listing potential exploiters of prostitution and we can observe the number of times an individual is charged. Obviously, we cannot observe the units not captured by the Justice system. We aim to estimate the hidden part of the population, i.e., the size of those unreported to the Public Prosecutor's offices. Capture-recapture models have already been used to investigate prostitution and sex workers; see, for instance, Rossmo & Routledge (1990), which estimates the number of street prostitutes in 1986/1987 in Vancouver, and Roberts Jr & Brewer (2006), which estimates the number of their clients. In this Chapter, we aim to estimate the size of prostitution exploiters, rather than the number of prostitutes or their clients. Our data on *prostitution exploiters* refer to perpetrators of adult sexual exploitation, according to the international classification ICCS (UNODC (2015)); these crimes include recruiting, enticing or procuring a person into prostitution; pimping; keeping, managing or knowingly financing a brothel; knowingly letting or renting a building or other place for the purpose of the prostitution of others.

Figure 2.3 depicts our data. The total number of observed prostitution exploiters is $n = 2740$, the "one" counts are $n_1 = 2269$. Counts greater than 5 are relatively few; 12 is the maximum number of observed captures.

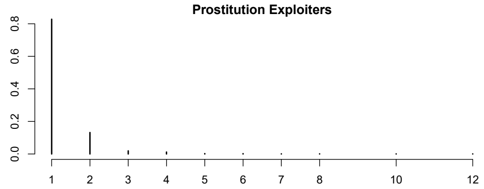


Figure 2.3. Relative frequencies of observed counts for prostitution exploitation data in Italy in 2014

We compared all 3 basic models analysed in this Chapter and their one–inflated counterparts on this data. In all one–inflated models we set a uniform $\omega \sim Beta(1, 1)$. We set $p \sim Beta(1, 1)$ in the Geometric and OIG models, and $\lambda \sim Gamma(0.01, 0.01)$ in the Poisson and OIP. Different values for the Gamma prior were also tested, obtaining very similar results. As for the Negative Binomial, the boundary problem emerged clearly, as, when adopting a $Gamma(0.1, 0.1)$ prior for r, we obtained a posterior mean for N twenty times greater than any other model (498000). For this reason, we opted for an $InvGamma(0.1, 0.1)$, both on the NB and the OINB models. In all cases, the number of replications of the MCMC algorithm is set to 10^6 with a thinning of 20 observations. As priors over N, we tried both Rissanen's and the improper $p(N) \propto 1/N$. The two alternatives gave almost identical results. Standard diagnostic tools confirmed the convergence of the algorithms.

The results are summarized in Table 2.5 and in Figure 2.4. Figure 2.4 shows the estimated posterior distributions of n_0 and of the parameters of the one–inflated models. The regular shape of the posterior distributions is evident from Figure 2.4, so the differences in adopting the posterior mode, median or mean are quite negligible. Regularity of the posterior distributions was consistently observed in all the applications and simulations

Table 2.5. The posterior mode and credible intervals for the population size N, posterior mean for ω and model parameters for prostitution exploitation data

Estimator/Model	$\hat{N}$	95%CI.$\hat{N}$	$\hat{\lambda}$	p	r	
Ignoring one–inflation						
Poi	7210	6780 - 7689	0.476			
Geo	13332	12415 - 14394			0.795	
NB	89140	35162 - 188368		0.665	0.088	
Chao	9851	8961 - 10868				
Zelterman	10030	9033 - 11027	0.319			
Modeling one–inflation						$\hat{\omega}$
OIP	3895	3656 - 4156	1.213			0.645
OIG	8182	7406 - 9233		0.669		0.478
OINB	19566	6174 - 71710		0.580	0.213	0.363
Mod.Chao.OIP	6493	4163 - 8823				
Mod.Chao.OIG	19628	9143 - 30112				

presented in this Chapter. Regularity of the posterior distributions does not hold for the n_0 and the r of the OINB model, due to the boundary problem.

In the upper part of Table 2.5 we give the estimates deriving from the Poisson, Geometric and Negative Binomial that ignore one–inflation and compare them to the well-known Chao and Zelterman estimators (see Chao (2014) for a detailed description). In the lower part of the Table, we give the results from the one–inflated counterparts of the 3 models and compare them to the modified Chao estimators, as suggested in Böhning et al. (2018). This estimator depends on the baseline distribution; we evaluate it assuming both Poisson and Geometric distribution with one–inflation (Mod.Chao.OIP and Mod.Chao.OIG, respectively), as in Böhning & Ogden (2021).

In Figure 2.3, the presence of one–inflation seems likely, and is, in fact, largely confirmed by the test introduced in Section 2.2.3. Both the OIP and the OIG have posterior probabilities several orders of magnitudes greater than the Poisson and the Geometric. The log marginal likelihoods are: -1863.39 (Poi), -1756.23 (Geo), -1718.21 (OIG), -1761.95 (OIP). The OINB model was found to have by far the highest log marginal likelihood, namely -1712.25. However, we believe that caution should be used in adopting the estimates from the OINB. In fact, the boundary problem seems evident ($\hat{r} = 0.2$), and the uncertainty contained in the estimate of n_0 is excessive (the width of the interval estimates is about 25 times greater than the total number of observed units).

As expected, if we ignore one–inflation, we risk severely overestimating the population size. Geometric and Negative Binomial distributions account for heterogeneity and produce much larger estimates than the Poisson distribution.

2.5.1 Results from some popular case-studies

In this Section, we apply the Bayesian model to a selection of well-known cases popular in the capture–recapture literature. We consider the following real cases:

1 street prostitutes in Vancouver: the data show the count of prostitution arrests made by the Vancouver Police Department Vice Squad for engaging in prostitution in 1986/1987, initially presented and analysed by Rossmo & Routledge (1990);

2 opiate users in Rotterdam: the data show the number of applications for a methadone treatment program made by opiate users in Rotterdam in 1994, first reported and

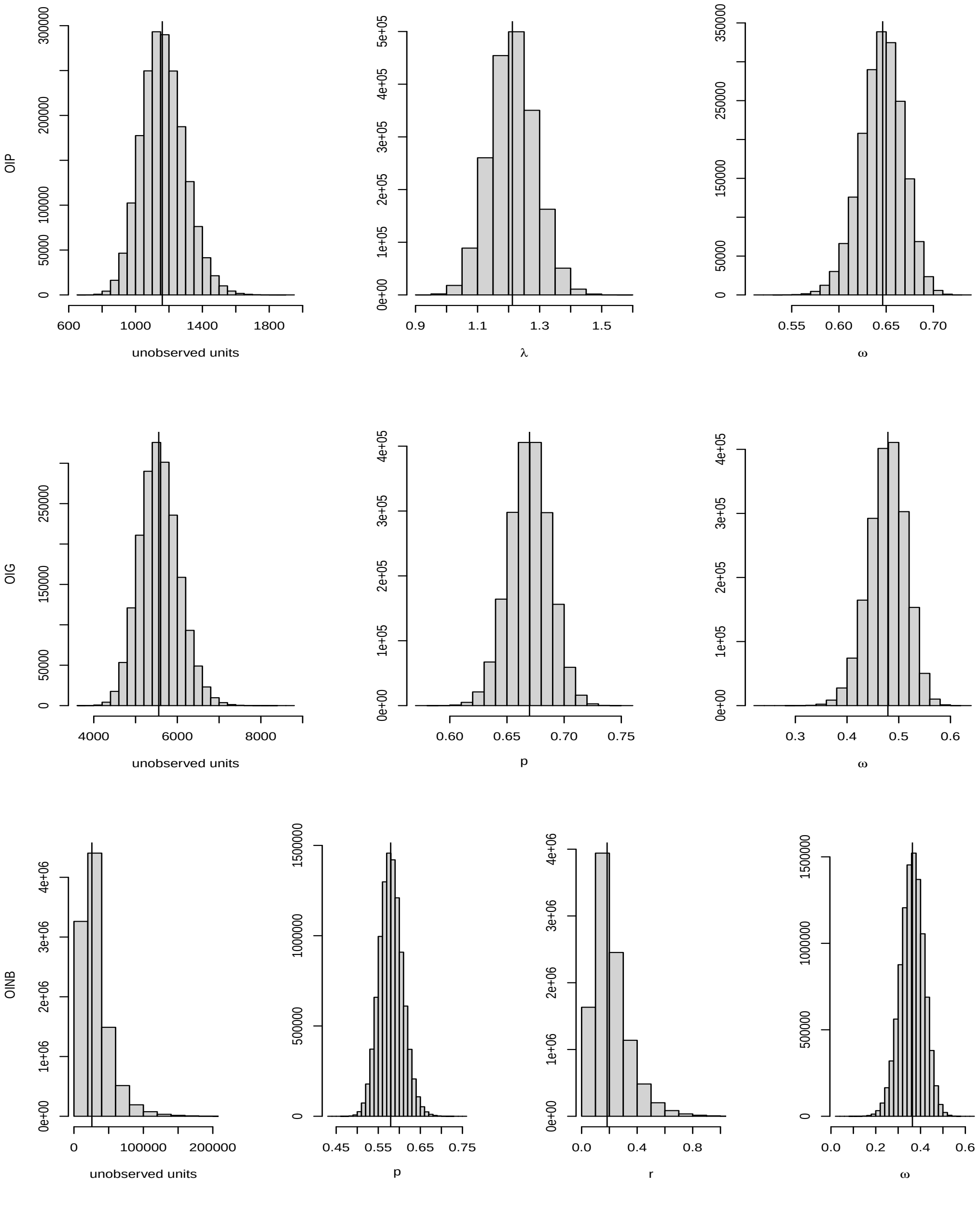


Figure 2.4. Posterior distributions of n_0 and of the parameters of all one–inflated models for prostitution exploitation data. Vertical lines show the posterior medians.

Table 2.6. Observed count distribution for three real cases

Real Cases	Counts										
1. Prostitutes	n_1	n_2	n_3	n_4	n_5	n_6	n				
	541	169	95	37	21	23	886				
2. Opiate users	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	n
	1206	474	198	95	29	19	5	2	0	1	2029
3. Heroin users	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	n_{11}
	2176	1600	1278	976	748	570	455	368	281	254	188
	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	n_{17}	n_{18}	n_{19}	n_{20}	n_{21}	n
	138	99	67	44	34	17	3	3	2	1	9302

analysed by Cruyff & van der Heijden (2008);

3 heroin users in Bangkok: the data provide the counts of treatment episodes by heroin users in Bangkok in 2002, available in Viwatwongkasem et al. (2008) and previously analysed by Böhning et al. (2004).

The observed count distribution of the three real cases are shown in Table 2.6. In the Vancouver prostitutes dataset, we observe $n = 886$ individuals and the number of units captured once is $n_1 = 541$. The Rotterdam opiate–user dataset contains $n = 2029$ units and $n_1 = 1206$. The Bangkok heroin–user dataset provides $n = 9302$ observations with $n_1 = 2176$.

These data sets have been widely examined in capture–recapture literature, also under the one–inflation hypothesis (see Godwin & Böhning (2017) and Godwin (2017)).

We apply our models to the three case–studies, with the following prior settings: For the Poisson and OIP models we set, a priori, $\omega \sim Beta(1,1)$ and $\lambda \sim Gamma(0.1, 0.1)$. In the OINB model we set $r \sim InvGamma(0.1, 0.1)$ and $p \sim Beta(1,1)$. In all our applications, the number of replications of the MCMC algorithm is 10^6 with a thinning of 20 observations. Standard diagnostic tools confirmed the convergence of the algorithm. The results for all three datasets are summarized in Table 2.7, which shows the posterior modes and credible intervals of N, and the posterior means of the model parameters.

The presence of one–inflation in these datasets is less severe than in the prostitution exploitation data analysed in the previous Section. However, as expected, estimates from the base distributions are consistently greater than the corresponding one–inflated estimates, confirming that we might be overestimating the population size if we ignore one–inflation.

For the Vancouver prostitute data, our model selection strategy strongly suggests the OINB distribution, its posterior probability being several orders of magnitudes greater than the competing models. The inflation rate ω is estimated around 0.40. The base Negative Binomial encounters the boundary problem, as is clear from the r estimate and even more from the credible intervals for N. OINB and OIP models produce similar estimates for N, with the credible intervals mostly overlapping (the 95%HPD under OINB is slightly greater than under OIP), whilst the OIG's credible interval barely overlaps the others.

As for the Rotterdam opiate–user data, Bayesian model selection largely favours the Geometric distribution, with a posterior probability of 0.89, against 0.104 and 0.006 for OIG and OINB, respectively; the Poisson models posterior probabilities being negligible, both the baseline and the one–inflated. In this case, the one–inflation does not seem to affect the data.

The posterior model probabilities for Bangkok heroin–user data favor the OINB model, even though the estimated inflation rate is quite low, a mere 0.056. The boundary problem is not an issue with this dataset, since the estimate of r is rather greater than 1.

Table 2.7. The posterior mode and credible intervals for the population size N, posterior mean for ω and model parameters, for real cases

1. Prostitutes in Vancouver		$\hat{N}$	95%HPD($\hat{N}$)	$\hat{\omega}$	$\hat{\lambda}$	$\hat{r}$	$\hat{p}$
Model	Poi	1240	1177 – 1300		1.254		
	Geo	2045	1906 – 2217				0.570
	NB	3340	1977 – 167925			0.145	0.395
	OIP	1017	982 – 1058	0.438	2.037		
	OIG	1820	1669 – 2003	0.192			0.517
	OINB	1040	991 – 1238	0.399		19.104	0.862
	Mod.Chao.OIP	1005	933 – 1077				
	Mod.Chao.OIG	1421	1097 – 1745				
2. Opiate users in Rotterdam		$\hat{N}$	95%HPD($\hat{N}$)	$\hat{\omega}$	$\hat{\lambda}$	$\hat{r}$	$\hat{p}$
Model	Poi	2934	2832 – 3038		1.174		
	Geo	4913	4676 – 5188				0.588
	NB	4960	4244 – 6818			0.869	0.566
	OIP	2500	2418 – 2587	0.336	1.663		
	OIG	4796	4491 – 5085	0.047			0.577
	OINB	3213	2616 – 4665	0.157		2.861	0.692
	Mod.Chao.OIP	2633	2398 – 2867				
	Mod.Chao.OIG	4745	3691 – 5799				
3. Heroin users in Bangkok		$\hat{N}$	95%HPD($\hat{N}$)	$\hat{\omega}$	$\hat{\lambda}$	$\hat{r}$	$\hat{p}$
Model	Poi	9452	9427 – 9477		4.134		
	Geo	12206	12064 – 12341				0.238
	NB	11572	11357 – 11817			1.232	0.267
	OIP	9364	9349 – 9380	0.207	5.004		
	OIG	12195	12056 – 12334	0.003			0.237
	OINB	10826	10606 – 11098	0.056		1.627	0.302
	Mod.Chao.OIP	9859	9757 – 9961				
	Mod.Chao.OIG	11810	11350 – 12270				

In all cases, the OINB model produces estimates for N higher than the OIP and lower than OIG. Also the one–inflation rate estimates under the OINB model prove always lower than the estimates obtained from the OIP model and higher than those from the OIG. It appears that by using the OINB, part of the one-inflation component identified by the OIP is instead explained through the two parameters of the Negative Binomial. The credible intervals of the OIP are consistently smaller than those of the competing models, and barely overlap, with the exception of Vancouver prostitute data, where actually the OINB model tends to the OIP one (note the high estimates for the parameter r).

The results in Table 2.7 can be compared with non Bayesian results reported in Godwin & Böhning (2017) and Godwin (2017), for the one–inflated Poisson and Negative Binomial models. We note that the use of weakly informative priors leads to results that are close to the frequentist approach. Moreover, the results from our Bayesian model selection strategy are also confirmed by likelihood ratio tests proposed in Godwin (2017), even if likelihood ratio tests provide less strong evidence than our results.

2.6 Concluding remarks and future works

In this Chapter we have dealt with the issue of one–inflation on repeated count data in population size estimation, adopting a fully Bayesian approach. We discussed our model for one–inflation under an unspecified count distribution, describing a general Gibbs sampler. Specifically, we derived the conditional distributions of the model parameters under the Poisson and Geometric assumption; moreover, to deal with data that show over–dispersion, we also illustrated the Bayesian analysis for the Negative Binomial model. We considered the boundary problem of the Negative Binomial distribution; in the Bayesian setting the prior parameter specification might help alleviate it. A fully Bayesian model selection approach, which includes testing for the one–inflation assumption, was developed for all the distributions considered in the Chapter.

Alongside the usual advantages of a Bayesian approach, namely the possibility of incorporating any prior knowledge in the analysis and ease in producing interval estimates of any quantity as a by-product of the estimation procedure, we recognize a less obvious point in favour. In fact, although, admittedly, it is not common to have prior information on the quantities at hand, even weakly informative priors can have a positive impact on the analysis. As we saw in Section 2.4.3, the use of a weakly informative prior when using a Negative Binomial model or its one–inflated counterpart can help stabilize the estimation procedure and avoid the “boundary problem” in case of moderate severity. On the other hand, the choice of the prior distribution for the size parameter of the Negative Binomial may affect model selection procedures which require additional investigation in order to allow a more general use of such distribution in capture recapture models.

We are currently working on extensions of the current model to cope with observed and unobserved heterogeneity in the presence of one-inflation, exploiting individual covariates, and introducing more complex hierarchical structures and mixing models.

Moreover, we are considering the possibility of taking model uncertainty into account with a model averaging technique in a single procedure by exploiting the reversible jump algorithm (see Green (1995)).

In addition, when dealing with sensible data, like the prostitution exploitation data which do not share a unique identifier, we may encounter record linkage problems. In this case, it would be important also to take into account the record linkage process uncertainty in population size estimation; see Tancredi & Liseo (2011). Note also that linkage errors can themselves produce one-inflation. In fact, when matching information does not suffice to recognise multiple captures of the same individual, the resulting missing links erroneously increase the number of singletons. However, it is worth nothing that, unlike the case with the framework considered in this Chapter, linkage errors also affect the observed sample size n.

Finally, we are investigating more general behavioural mechanisms producing different forms of inflation. For example, we could assume that when the latent count y^* is equal to k, instead of necessarily having an observation y equal to 1 or to the true value k, we have that y follows a mixture of two distributions. In particular we may have a mixture component with weight $1 - \omega$ concentrated on the latent value $y^* = k$. The other component with weight ω may have support on the set $\{1, \dots, k\}$ and can, for example, be a Binomial(k, ψ) truncated on 0. Thus, when $\psi = 0$ we have exactly the form of inflation discussed in this Chapter while when $\psi > 0$ the model also allows us to inflate counts greater than one, generalizing the effects of the behavioural mechanism.

Appendix *(Marginal likelihood calculations)*

Expression (2.4) for the marginal likelihood is obtained by observing that

$$\begin{aligned}
p(\mathbf{y}|M_i) &= \int \sum_{N=n}^{\infty} f(\mathbf{y}|\theta_i, N, M_i) p(N) p(\theta_i) d\theta_i \\
&= \int \sum_{N=n}^{\infty} \binom{N}{n} f(0|\theta_i)^{N-n} \prod_{i=1}^{n} f(y_i|\theta_i) \frac{c}{N} p(\theta_i)\, d\theta_i \\
&= c \int \sum_{n_0=0}^{\infty} \frac{(n+n_0)!}{n!\, n_0!} \frac{1}{n+n_0} f(0|\theta)^{n_0} (1 - f(0|\theta_i))^n \prod_{i=1}^{n} \frac{f(y_i|\theta_i)}{1 - f(0|\theta_i)} p(\theta_i) d\theta_i \\
&= \frac{c}{n} \int \sum_{n_0=0}^{\infty} \binom{n+n_0-1}{n-1} f(0|\theta)^{n_0} (1 - f(0|\theta_i))^n \prod_{i=1}^{n} \frac{f(y_i|\theta_i)}{1 - f(0|\theta_i)} p(\theta_i) d\theta_i \\
&= \frac{c}{n} \int \prod_{i=1}^{n} \frac{f(y_i|\theta_i)}{1 - f(0|\theta_i)} p(\theta_i) d\theta_i .
\end{aligned}$$

Chib's approximation is based on the identity

$$p(\mathbf{y}|M_i) = \frac{f(\mathbf{y}|\theta_i, N) p(\theta_i) p(N)}{p(\theta_i, N|\mathbf{y}, M_i)}$$

valid for each point (θ_i, N). To approximate the marginal likelihood we may select a point $(\tilde{\theta}_i, \tilde{N})$ given, for example, by the posterior means obtained with a first run of the Gibbs sampler and then estimate the value of the posterior $p(\tilde{\theta}_i, \tilde{N}|\mathbf{y}, M_i)$ via a second run by using the following strategies.

For the Poisson model M_i, where $\theta_i = \lambda$, suppressing the model dependence in the notation hereafter, we have $p(\tilde{\theta}, \tilde{N}|\mathbf{y}) = p(\tilde{\lambda}, \tilde{N}|\mathbf{y}) = p(\tilde{N}|\tilde{\lambda}, \mathbf{y}) p(\tilde{\lambda}|\mathbf{y})$ and the only quantity that need to be estimated is $p(\tilde{\lambda}|\mathbf{y})$. Anyway

$$p(\tilde{\lambda}|\mathbf{y}) = \sum_{N} p(\tilde{\lambda}, N|\mathbf{y}) = \sum_{N} p(\tilde{\lambda}|\mathbf{y}, N) p(N|\mathbf{y})$$

and by exploiting the T realizations $N_{(1)} \dots, N_{(T)}$ of $p(N|\mathbf{y})$ from a second run of the Gibbs sampler we can estimate $p(\tilde{\lambda}|\mathbf{y})$ by

$$p(\tilde{\lambda}|\mathbf{y}) \approx \frac{1}{T} \sum_{t=1}^{T} p(\tilde{\lambda}|\mathbf{y}, N_{(t)})$$

where $p(\tilde{\lambda}|\mathbf{y}, N_{(t)})$ is the density of a Gamma$(\alpha_\lambda + s, \beta_\lambda + N_{(t)})$.

Similarly, for the Geometric model, where $\theta = p$, we have $p(\tilde{\theta}_i, \tilde{N}|\mathbf{y}) = p(\tilde{p}, \tilde{N}|\mathbf{y}) = p(\tilde{N}|\tilde{p}, \mathbf{y})p(\tilde{p}|\mathbf{y})$ and the only quantity that need to be estimated is $p(\tilde{p}|\mathbf{y})$. Anyway

$$p(\tilde{p}|\mathbf{y}) = \sum_N p(\tilde{p}, N|\mathbf{y}) = \sum_N p(\tilde{p}|\mathbf{y}, N)p(N|\mathbf{y})$$

and by exploiting the T realizations $N_{(1)} \ldots, N_{(T)}$ of $p(N|\mathbf{y})$ from a second run of the Gibbs sampler we can estimate $p(\tilde{p}|\mathbf{y})$ by

$$p(\tilde{p}|\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} p(\tilde{p}|\mathbf{y}, N_{(t)})$$

where $p(\tilde{p}|\mathbf{y}, N_{(t)})$is the density of a Beta $(\alpha_p + N_{(t)}, \beta_p + s)$.

For the OIP model where $\theta = (\lambda, \omega)$ we have $p(\tilde{\theta}, \tilde{N}|\mathbf{y}) = p(\tilde{\lambda}, \tilde{\omega}, \tilde{N}|\mathbf{y}) = p(\tilde{N}|\tilde{\lambda}, \tilde{\omega}, \mathbf{y})p(\tilde{\lambda}, \tilde{\omega}|\mathbf{y})$. In this case we need to estimate $p(\tilde{\lambda}, \tilde{\omega}|\mathbf{y})$ where

$$p(\tilde{\lambda}, \tilde{\omega}|\mathbf{y}) = \sum_N \sum_{\mathbf{y}^*} p(\tilde{\lambda}, \tilde{\omega}, N, \mathbf{y}^*|\mathbf{y}) = \sum_N \sum_{\mathbf{y}^*} p(\tilde{\lambda}, \tilde{\omega}|\mathbf{y}, N, \mathbf{y}^*)p(N, \mathbf{y}^*|\mathbf{y}).$$

Then, by exploiting the T realizations $\mathbf{y}^*_{(1)}, N_{(1)}, \ldots, \mathbf{y}^*_{(T)}, N_{(T)}$ of $p(\mathbf{y}^*, N|\mathbf{y})$ from the the first Gibbs sampler run, we can estimate $p(\tilde{\lambda}, \tilde{\omega}|\mathbf{y})$ by

$$p(\tilde{\lambda}, \tilde{\omega}|\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} p(\tilde{\lambda}, \tilde{\omega}|\mathbf{y}, \mathbf{y}^*_{(t)}, N_{(t)}).$$

Note that λ and ω are conditionally independent given $\mathbf{y}, \mathbf{y}^*$ and N. Moreover the conditional distribution $\lambda|\mathbf{y}, \mathbf{y}^*, N$ is Gamma$(\alpha_l + \sum_{k>0} kn^*_k, \beta_l + N)$ while the conditional distribution $\omega|\mathbf{y}, \mathbf{y}^*, N$ is Beta$(\alpha_\omega + n_z, \beta_\omega + \sum_{k>1} n_k)$.

Similarly, for the OIG model where $\theta = (p, \omega)$ we can follow exactly the same strategy by factorizing the posterior distribution as $p(\tilde{\theta}, \tilde{N}|\mathbf{y}) = p(\tilde{p}, \tilde{\omega}, \tilde{N}|\mathbf{y}) = p(\tilde{N}|\tilde{p}, \tilde{\omega}, \mathbf{y})p(\tilde{p}, \tilde{\omega}|\mathbf{y})$ and estimating $p(\tilde{p}, \tilde{\omega}|\mathbf{y})$ by

$$p(\tilde{p}, \tilde{\omega}|\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} p(\tilde{p}, \tilde{\omega}|\mathbf{y}, \mathbf{y}^*_{(t)}, N_{(t)}).$$

where $\mathbf{y}^*_{(1)}, N_{(1)}, \ldots, \mathbf{y}^*_{(T)}, N_{(T)}$ are T realizations from $p(\mathbf{y}^*, N|\mathbf{y})$ obtained from the first Gibbs sampler run. Also in this case $\tilde{p}$ and $\tilde{\omega}$ are conditionally independent given $\mathbf{y}, \mathbf{y}^*, N$. The conditional distribution $p|\mathbf{y}, \mathbf{y}^*, N$ is Beta$(\alpha_p + N, \beta_p + \sum_{k>0} n^*_k)$ while and $\omega|\mathbf{y}, \mathbf{y}^*, N$ is Beta$(\alpha_\omega + n_z, \beta_\omega + \sum_{k>1} n_k)$.

For the Negative Binomial model we have $\theta = (p, r)$ and the posterior can be factorized as

$$p(\tilde{\theta}, \tilde{N}|\mathbf{y}) \quad = \quad p(\tilde{p}, \tilde{r}, \tilde{N}|\mathbf{y}) = p(\tilde{N}|\tilde{p}, \tilde{r}, \mathbf{y})p(\tilde{p}|\tilde{r}, \mathbf{y})p(\tilde{r}|\mathbf{y}).$$

where, as in the previous models, the conditional density $p(\tilde{N}|\tilde{p}, \tilde{r}, \mathbf{y})$ is known. The conditional density $p(\tilde{p}|\tilde{r}, \mathbf{y})$ can be obtained by an extra run of the Gibbs sampler with r fixed to $\tilde{r}$. In fact

$$p(\tilde{p}|\tilde{r}, \mathbf{y}) = \sum_N p(\tilde{p}, N|\tilde{r}, \mathbf{y}) = \sum_N p(\tilde{p}|N, \tilde{r}, \mathbf{y})p(N|\tilde{r}, \mathbf{y})$$

and the conditional distribution $p|N,\tilde{r},\mathbf{y}$ is Beta with parameters α_p+Nr,β_p+s. Instead the calculation of the marginal posterior $p(\tilde{r}|\mathbf{y})$ can be obtained following the approach proposed by Chib & Jeliazkov (2001).

For the OI Negative Binomial we have $\theta=(p,r,\omega)$ and the posterior can be factorized as

$$p(\tilde{\theta},\tilde{N}|\mathbf{y}) \quad = \quad p(\tilde{p},\tilde{\omega},\tilde{r},\tilde{N}|\mathbf{y}) = p(\tilde{N}|\tilde{p},\tilde{r},\tilde{\omega},\mathbf{y})p(\tilde{\omega},\tilde{p}|\tilde{r},\mathbf{y})p(\tilde{r}|\mathbf{y}).$$

Also in this case the conditional density $p(\tilde{N}|\tilde{\omega},\tilde{p},\tilde{r},\mathbf{y})$ is known and $p(\tilde{\omega},\tilde{p}|\tilde{r},\mathbf{y})$ can be obtained by an extra run of the Gibbs sampler with r fixed to $\tilde{r}$ by

$$p(\tilde{p},\tilde{\omega}|\tilde{r},\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} p(\tilde{p},\tilde{\omega}|\tilde{r},\mathbf{y},\mathbf{y}^*_{(t)},N_{(t)}).$$

Note that the parameters p and ω are conditionally independent given $r,\mathbf{y},\mathbf{y}^*$ and N with $p|r,\mathbf{y},\mathbf{y}^*$ and N which is Beta$(\alpha_p+Nr,\beta_p+\sum_{k>0}kn^*_k)$ and $\omega|\mathbf{y},\mathbf{y}^*,N$ which is, as in the previous inflated models, Beta$(\alpha_\omega+n_z,\beta_\omega+\sum_{k>1}n_k)$. Finally, as for the non–inflated Negative Binomial counterpart, the calculation of the marginal posterior $p(\tilde{r}|\mathbf{y})$ can be obtained following the approach proposed by Chib & Jeliazkov (2001).

Chapter 3

Semi-parametric Bayesian approach for estimating the size of criminal populations modeling the excess of singletons

Abstract

In this Chapter we aim at estimating the size of certain criminal populations on the basis of administrative data on judicial proceedings by exploiting capture–recapture models for repeated count data. The data at our disposal exhibit an abundance of units that are captured exactly once, which suggests the necessity of explicitly modeling this deviation. We distinguish two possible causes for this phenomenon, namely, the erroneous inclusion of out–of–scope units, and a particular behavioral effect preventing subsequent captures after the first one. Accordingly, we propose two families of one–inflated models to estimate the number of uncaptured units. We propose a Bayesian semi-parametric approach by considering a Dirichlet process mixture model as a base model, and extend this class to include one–inflation. The proposed model and the two one–inflated counterparts are compared on three datasets of Italian criminal proceedings.

3.1 Introduction

The need to estimate the number of people involved in a certain illegal activity is driven by several factors, including social, judicial and economic ones. Such an estimate, in fact, would allows us to better understand the size of the illegal phenomenon itself, and to assess the threat it poses to society; it serves to better size the police forces to counter it and to evaluate the effectiveness of prevention and counteraction policies. In addition to social and criminology purposes, some illegal activities have also a great economic interest. Indeed, the European Parliament and Council have identified the smuggling of goods, prostitution exploitation and drug trafficking as the main sources of illegal economic transactions to report in the national accounts aggregates and in the GDPs of the member states (Regulation EU No 549/2013 of the European Parliament and of the Council (ESA 2010)[1], based on the international recommendation on System of National Accounts (SNA) 2008). In this Chapter we aim to estimate the number of people involved in these three illegal activities in Italy during 2014.

In Italy, smuggling activities mainly regards cigarettes. The internal tobacco market is regulated, and smuggling is related to three product types: (1) original cigarettes manufactured by legitimate business enterprises imported beyond the limit or through an illegal supply chain; (2) the so called “cheap white”, i.e. cigarettes manufactured by legitimate business enterprises with a large share of the production being sold without all applicable duties paid, usually outside the jurisdiction where they are produced since there is not enough internal demand in those countries; (3) counterfeit cigarettes, which bear brands without the owner’s permission and which are often produced in countries characterized by low labor costs and by the presence of strong economies of scale in tobacco processing. Contraband cigarettes arrive in Italy especially from Eastern European countries, China and the United Arab Emirates. The internal production is considered negligible, as well as the exportation.

Prostitution is neither prosecuted nor regulated in Italy, while the trafficking, exploitation, and aiding and abetting of prostitution is a crime, regulated by law and prosecuted. This activity is mostly under the control of organized crime. For the estimate of national account aggregates, prostitution is considered as a service to the households and the number of prostitutes are estimated without considering if they are in a condition of “exploitation”. On the other side, prostitution exploitation is also related to human trafficking, since the sexual exploitation and forced labor are the main forms of exploitation for trafficking in persons. To this regard, the United Nations Protocol to Prevent, Suppress and Punish Trafficking

[1] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32013R0549

in Persons, the so-called Palermo Protocol, which provides the basis for much national legislation, stipulates that the fight against human trafficking must be equally focused on the arrest and punishment of offenders and providing protection and assistance to victims, with full respect for their human rights. Hence, the need of a more comprehensive estimation of the true volume of presumed prostitution exploiters, so more targeted policy interventions can be implemented to improve the detection of – and prevention to – the most hidden populations.

Italy is considered as a key transit area for global drug trafficking routes, both for the central position in the Mediterranean Sea and in Europe, for the presence of criminal organizations, and for the relevant consumer markets. The drug market mainly concerns importation and exportation, rather than production. Cocaine, cannabis, and amphetamine-type stimulants are the main substances in transit, managed by national and international organized criminal networks. For this reason, the fight against drug trafficking coincides, in most cases, with the fight against mafia-type organizations. For these criminal associations, drug trafficking remains the "main multiplier of wealth", since its profits are by far the most significant of those generated by any other human activity, both lawful and illegal. Therefore, the estimates of drug traffickers remains decisive, both because it is essential to contain the spread of drugs, which affects health and public order, and because it is essential to reduce the strength and the wealth of criminal organizations and of the whole complex chain that revolves around them.

People involved in illegal activities have obvious reasons to hide their businesses to the public, the administrative system, the police and of course they also escape traditional statistical surveys. In this work we exploit administrative registers coming from the Ministry of Justice, which report alleged crimes for which the judicial authority has collected enough evidences to decide to start a criminal proceeding. These data are usually utilized by National Statistical Offices to produce official crime statistics. National Statistical Offices usually warn users that these statistics refer only to crimes recorded by the authorities, based on reports to the police from victims and witnesses. In this Chapter, we propose an additional re–use of these sources to estimate the hidden part of the population, those involved in criminal activities but for some reasons unreported to the justice system.

Crimes records in the registers of the Public Prosecutor's offices, contain soft identifiers of the denounced subjects, namely date and place of birth and gender. On the basis of this information, crime authors can be recognized within the register and followed in a specific time span. In this way, it is possible to count the number of times each unit appears in the Prosecutor's offices registers. Obviously, we do not observe units not caught by the Justice system. Hence, the registers can be considered as incomplete lists of potential criminals, since only denounced crimes and suspected criminals are reported. We aim to estimate the total criminal population size (reported or not) via capture–recapture methodologies.

The data at our disposal are not provided with information on the exact date of each report. Luckily, the assumption of homogeneous capture probabilities for each unit throughout the reference year (i.e., time–homogeneity) seems reasonable. Under time–homogeneity, in absence of individual covariates, the data can be simply summarized as counts of units captured j times, $j = 1, 2, ...$, commonly called "repeated count data" in the context of capture–recapture literature. The common parametric approach to analyze this data is to define a counting distribution for the number of captures in the population. In absence of any additional individual information, it is crucial to model the unobserved heterogeneity in the captures probabilities. A common approach to this end is represented by the use of mixtures of counting distributions which is well established in capture–recapture, see, e.g., Norris & Pollock (1996), Pledger et al. (2003), Böhning et al. (2005). In this work we propose a Bayesian approach based on mixtures of Poisson distributions.

The choice of the number of components in a finite mixture model is a long–debated problem in model selection. We address this aspect in a fully Bayesian approach by resorting to a semi–parametric modeling to avoid a definite choice over that number, by considering a Dirichlet process mixture (DPM) model. There are precedents for the use of DPMs in capture–recapture literature: a DPM of Poisson distributions was proposed in Guindani et al. (2014) for modeling gene expression sequence abundance, and, in the context of multiple systems estimation, a DPM approach to latent class models was proposed in Manrique-Vallier (2016) to estimate civilian casualties in war.

We also compare the DPM models with sparse finite mixture (SFM) models which, to the best of our knowledge, even if strictly related to DPMs, have not yet been applied in this field. For a completely different semi–parametric Bayesian approach to mixture of Poisson distribution in capture–recapture see Fegatelli & Tardella (2018).

The data at our disposal (presented in the next Section) exhibit an elevated number of individuals captured exactly once (sometimes referred to as "singletons"). The excess of singletons has been studied in several works in the recent literature in capture–recapture, where it is known as "one–inflation" (see, e.g., Godwin & Böhning (2017), Godwin (2017), Bunge et al. (2014), 2). The idea is that there exists a mechanism which increases the number of observed singletons with respect to a baseline counting distribution. Ignoring an existing mechanism of one–inflation can lead to a severe overestimation of the population size, so we included in our analysis the one–inflated counterpart of a DPM of Poissons (our baseline distribution). We identified three possible sources of one–inflation and derived a one–inflated model for two of them.

The three possible one–inflation mechanisms are presented and commented with respect to the criminal data in Section 3.2. The DPM of Poisson models are presented in Section 3.3 together with a Gibbs–based MCMC for the estimation of the posterior distribution of the population size. Two one–inflated mixtures of Poisson models based on two different one–inflation mechanisms are presented in Section 3.4, together with the relative Bayesian estimation algorithms. In Section 3.6 we present the results of all our models for estimating the three criminal populations. Some remarks conclude the Chapter in Section 3.7. In the Supplementary material we present the SFM models showing a comparison with the DPM approach.

3.2 Data on criminal activities and one–inflation

Figure 3.1 depicts the relative frequencies of observed number of captures for the individuals involved in the three criminal activities mentioned in the introduction in Italy during 2014. The total number of distinct individuals charged with smuggling is 3349. The top panel of Figure 3.1 shows the relative distribution by number of captures per person, which presents a maximum of 27 captures. The total number of distinct individuals facing charges of prostitution exploiting (middle panel) is 2740. Individuals with more than 5 proceedings are relatively few, and 12 is the maximum number of observed captures. The total number of distinct individuals charged with drugs trafficking is 34964. The relative distribution (bottom panel) presents a rather long right tail which is rarely observed in social applications of capture–recapture: a few units are captured even more than 70 times in the reference year.

A clear common characteristic of the three observed distributions is the high number of individuals captured once. In particular, we have that about 78% of the observed distinct individuals in the smugglers data are captured once. That percentage is about 82% in the prostitution exploiters data, and 77% in the drug traffickers data. Explicitly modeling a process generating an excess of singletons, that is, a mechanism of one–inflation, is of particular importance in capture–recapture, as it typically implies a substantial difference in the population size estimates. In particular, a one–inflated model always implies a lower estimate of the population size with respect to its baseline counterpart.

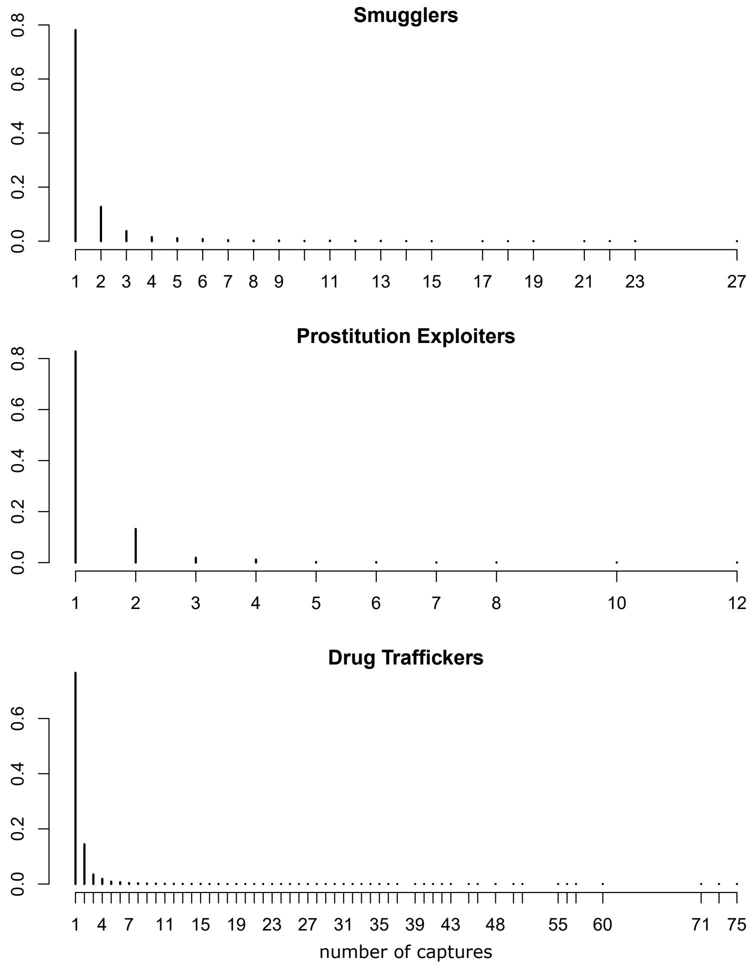


Figure 3.1. Relative frequency distribution of the counts of proceedings per person for smuggling (top), prostitution exploiting (mid), and drug trafficking (bottom) in Italy in 2014

One–inflation can be motivated by different factors, and we have identified the following list of possible generating mechanisms:

1. a specific behavior of the units, which learn how to avoid subsequent captures after the first one;

2. the presence of spurious units which do not belong to the reference population;

3. errors in re-identifying the units (linkage errors), due to the lack of unique identifiers and error-prone soft identifiers.

In our data we cannot exclude any of the previous factors having an impact on the observed excess of singletons. Clearly, people involved in illegal activities avoid any capture as much as possible. The behavior mechanism generating one–inflation in this case can be viewed as an extreme form of "trap shy" behavior. In animal abundance analysis, the term "trap shy" is often utilized to indicate a behavior where an animal, once captured, tends to stay away from traps, so that its probability of being captured is reduced at each additional capture. In our case, we can suppose that, after the first capture, a portion of persons involved in illegal activities acquires the necessary knowledge/ability to avoid any subsequent capture. This hypothesis has been explored in depths in several recent papers (see, e.g., Godwin & Böhning (2017), Godwin (2017), Godwin (2019), Böhning & Friedl (2021), Tajuddin et al. (2021), 2).

We cannot even rule out the second listed source of one–inflation, as our data might include spurious cases. Indeed, we analyze records from the reported offences for which the judicial authority started a proceeding, that is, according to the Italian law, after the reports of victims and witnesses, the investigative activities of the police authorities have collected sufficient evidence for the judicial authorities to take legal action. However, we actually do not have information on the outcome of the prosecutions, i.e., whether the person reported was sentenced in court or acquitted (and should not be included in our target population). A second possible source of spurious cases, peculiar for criminal records, has been mentioned in Tajuddin et al. (2021). It comes from the missclassification of the criminal offense, that is, individuals that are accidentally (or intentionally) charged with the wrong offense (see Nolan et al. (2011)). Clearly, limiting the spurious cases to the units captured once is a simplifying hypothesis, but it is undeniably reasonable. Then, we can hypothesize that a portion of the units captured once are out–of–scope, and should be discarded from the analysis. The presence of out–of–scope or spurious captures has been considered in the context of multiple systems estimation (MSE) by several authors (e.g., Overstall et al. (2014), Fegatelli et al. (2017), Di Cecco et al. (2018), Di Cecco (2019), Farcomeni (2020)). In MSE, data assume the form of a contingency table, and, in the above contributions, (a portion of) the cells are supposed to include units that should be discarded from the analysis, so that the observed counts represent an upper bound for the real values to be considered under the hypothesized models. In the context of repeated counts data, the hypothesis has been explored, with a different approach, for example in Bunge et al. (2014), Böhning et al. (2018) and Böhning & van der Heijden (2019) (more details will be given in subsequent sections). Finally, due to privacy issue, our data are not provided with a unique identifier, hence, we cannot exclude the possibility of one–inflation deriving from linkage errors. In fact, when matching information does not suffice to recognize multiple captures of the same individual, the resulting missing links erroneously increase the number of singletons. An accurate analysis of this hypothesis would require knowledge on the linkage procedure utilized as well as the availability of unique identifiers which we do not have access to. For this reason, we will focus on the first two sources of one–inflation and consequently provide two families of one–inflated models in Section 3.4. The interested reader can find additional information on linkage errors in MSE framework in Tancredi & Liseo (2011), Di Consiglio & Tuoto (2018), and Di Consiglio et al. (2019).

3.3 Dirichlet process mixtures

We assume that the population of criminals in a given year is closed of unknown size N. Let Y be the integer–valued random variable representing the number of times a given unit has been captured. We only observe the n individuals, $n \leq N$, which are captured at least once. Let n_j denote the number of units captured j times, such that $\sum_{j>0} n_j = n$. We want to estimate the number of uncaptured units n_0, or, equivalently, the total number of units in the population $N = n + n_0$. To take into account the heterogeneity of capture probabilities in the population, we assume that the number of times Y an individual appears in the Prosecutor's offices registers is a finite mixture of Poisson distributions. Denote as k the number of components of our mixture, as $f(j|\lambda_i)$ the probability $\lambda_i^j e^{-\lambda_i}/j!$ of being captured j times in the i–th Poisson component defined by the parameter λ_i. Let $\pi_1, \ldots, \pi_k$ be the mixing weights, such that $\sum_{i=1}^k \pi_i = 1$, and denote as θ the set of all parameters defining the model: $\{\pi_i\}_{i=1,\dots,k}$ and $\{\lambda_i\}_{i=1,\dots,k}$. Our model is then defined as:

$$P(Y=j) = f(j|\theta) = \sum_{i=1}^{k} \pi_i \; f(j|\lambda_i). \tag{3.1}$$

In order to avoid to specify in advance an unknown and fixed quantity for the number of mixture components, we adopt a semi-parametric approach by choosing a Dirichlet process mixture (DPM) prior. In particular, we assume the *truncated* version of the DPM, (see Ishwaran & James (2001)), where the weights $\pi_1, \dots, \pi_k$ of the components follow a finite stick–breaking process. In our DPM model, we fix an arbitrarily large value for k, and we set the following priors:

- a truncated stick–breaking process of parameter ϕ over the mixture weights, with a Gamma prior over ϕ:

  $$(\pi_1, ..., \pi_k) \sim SB(\phi), \qquad \phi \sim Gamma(\alpha_\phi, \beta_\phi);$$

- a conjugate Gamma prior for each parameter λ_i

  $$\lambda_i \sim Gamma(\alpha_\lambda, \beta_\lambda), \quad i = 1, ..., k,$$

  where β_λ can eventually have a hyper-prior Gamma distribution to permit a hierarchical prior modeling approach;

- an improper prior with distribution proportional to $1/N$ over the parameter N. Other weakly informative options for the total population size N are available in literature, (see, e.g., Wang et al. (2007) and Xu et al. (2014)). However, we found no significant difference in using them in our applications. In addition, assuming $P(N) \propto 1/N$ leads to a computational advantage, as it is well-known that in this case, we can directly sample from the full conditional distribution of $n_0 = N - n$ which is Negative Binomial.

Throughout the Chapter all Gamma distributions are to be intended as parameterized in terms of shape and rate parameters.

3.3.1 MCMC algorithm

Here we detail the Gibbs–based MCMC algorithm to sample from the posterior distribution of (N, θ). Let n_j^i be the (latent) number of units in the i–th component that have been captured j times. Let n^i be the total number of population units (captured or uncaptured) in component i: $n^i = \sum_{j \geq 0} n_j^i$. Then, at each iteration we have the following steps:

1. Update all parameters λ_i by sampling from their respective full conditionals:

   $$\lambda_i \sim Gamma\left(\alpha_{\lambda_i} + \sum_{j\geq 0} j \cdot n_j^i \,,\; \beta_{\lambda_i} + n^i\right), \quad i = 1, \dots, k.$$

2. In order to update all mixing weights π_i, sample

   $$V_i \sim Beta\left(1 + n^i, \phi + \sum_{h=i+1}^{k} n^h\right), \quad i = 1, \dots, k-1,$$

   then set $V_k = 1$ and

   $$\pi_i = V_i \prod_{h,h<i} (1 - V_h), \quad i = 1, \dots, k.$$

3. Sample ϕ from $Gamma(\alpha_\phi - 1 + k\,,\, \beta_\phi - \log \pi_k)$.

4. Sample n_0 from its full conditional, that, given the improper prior $P(N) \propto 1/N$, is

$$n_0 \sim NegBin\left(n, 1 - f(0|\theta)\right),$$

where the probability $f(0|\theta)$ of not being captured is calculated according to (3.1).

5. Update the number of units captured j times associated to each mixture component $(n_j^1, ..., n_j^k)$ by sampling from

$$Mult\left(n_j, (\rho_{1|j}, ..., \rho_{k|j})\right), \quad \text{for } j \geq 0,$$

where $\rho_{i|j}$ denotes the probability of belonging to the i–th component conditionally on the number of captures j and the current values of θ, and is calculated as:

$$\rho_{i|j} = \frac{\pi_i\, f(j|\lambda_i)}{\sum_{h=1}^{k} \pi_h\, f(j|\lambda_h)}.$$

3.4 One–inflated models

We consider two classes of one–inflated families defined by the first two listed sources of inflation mentioned in Section 3.2: the behavioral effect and the presence of spurious cases. Under the hypothesis of one–inflation caused by a specific behavioral effect, an individual that, without that effect, would face multiple captures, now has a positive probability ω of being captured just once. The hypothesis can be modeled as follows: let B be the latent indicator variable identifying the units having this behavior. Each individual has a marginal probability ω of belonging to this subpopulation for which $P(Y > 1|B = 1) = 0$. Denote as Y^* the latent number of captures of a given unit that we would observe in absence of the behavioral mechanism. Let $f^*(j|\theta) = P(Y^* = j \mid \theta)$ be the relative probability distribution, depending on a set of parameters θ. Then we have

$$P(Y = j|B = 0) = f^*(j|\theta) \text{ for all } j, \quad \text{and} \quad P(Y = j|B = 1) = \begin{cases} f^*(0|\theta) & \text{if } j = 0; \\ 1 - f^*(0|\theta) & \text{if } j = 1. \end{cases}$$

The resulting distribution for Y is the one–inflated model defined as:

$$P(Y = j \mid \theta, \omega) = f(j|\theta, \omega) = \begin{cases} f^*(0|\theta) & \text{if } j = 0; \\ (1 - \omega) f^*(1|\theta) + \omega(1 - f^*(0|\theta)) & \text{if } j = 1; \\ (1 - \omega) f^*(j|\theta) & \text{if } j > 1. \end{cases} \quad (3.2)$$

We will refer to f^* as the “baseline” distribution and to (3.2) as its one–inflated counterpart.

Under the hypothesis of one–inflation caused by the erroneous inclusion of out–of–scope units, we assume a constant probability of each unit captured exactly once of being spurious. Formally, if we denote as S the latent indicator variable identifying the spurious units, each unit has a marginal probability ψ of belonging to this subpopulation for which

$$P(Y = j \mid S = 1) = \begin{cases} 1 & \text{if } j = 1; \\ 0 & \text{otherwise.} \end{cases}$$

If we denote again as $f^*(j|\theta) = P(Y^* = j \mid \theta)$ the baseline probability distribution of the number of captures without one–inflation, the resulting distribution of Y is a mixture of f^* and a Dirac's measure over the value one:

$$P(Y = j \mid \theta, \psi) = f(j|\theta, \psi) = \begin{cases} (1-\psi)f^*(0|\theta) & \text{if } j = 0; \\ (1-\psi)f^*(1|\theta) + \psi & \text{if } j = 1; \\ (1-\psi)f^*(j|\theta) & \text{if } j > 1. \end{cases} \tag{3.3}$$

Models (3.2) and (3.3) represent two distinct hypotheses on the source of one–inflation, and lead to different estimates of n_0. Note however that, in our analysis, data are truncated in zero, so that they assume respectively the form:

$$P(Y = j \mid \theta, \omega, Y > 0) = \begin{cases} \dfrac{(1-\omega)f^*(1|\theta) + \omega(1 - f^*(0|\theta))}{1 - f^*(0|\theta)} & \text{if } j = 1; \\ \dfrac{(1-\omega)f^*(j|\theta)}{1 - f^*(0|\theta)} & \text{if } j > 1. \end{cases}$$

and

$$P(Y = j \mid \theta, \psi, Y > 0) = \begin{cases} \dfrac{(1-\psi)f^*(1|\theta) + \psi}{1 - (1-\psi)f^*(0|\theta)} & \text{if } j = 1; \\ \dfrac{(1-\psi)f^*(j|\theta)}{1 - (1-\psi)f^*(0|\theta)} & \text{if } j > 1. \end{cases}$$

Then, it can be showed with simple algebra that the two zero-truncated one–inflated families are equivalent for fixed θ, according to the following reparameterization:

$$\psi = 1 - \frac{1-\omega}{1 - \omega f^*(0|\theta)}. \tag{3.4}$$

This implies that the two possible sources of one–inflation cannot be distinguished on the basis of the likelihood in case of capture–recapture data. That is, even under identifiability within each one–inflated class of distributions, we do not have identifiability between those two classes (on this subject see, e.g., Link (2003) and Link (2006)). Note that nonidentifiability holds for the truncated distributions, and the two options would lead to different estimates of n_0 (in particular, ceteris paribus, the estimate of n_0 under (3.3) would always be smaller than that deriving from (3.2)). This fact does not represent a problem in those situations where we can opt for a definite choice between the two forms of one–inflation (for example, whenever the capturing mechanism cannot be affected by units behavioral effects). As we have said before, in our analysis we are bounded to consider both options and appreciate the difference in the estimates under the two different assumptions. A formal approach in a Bayesian analysis would be that of considering model averaging, that is, the posterior of N averaged over the two models. In fact, even if the likelihoods are identical, the use of informative priors would lead us to different posterior probabilities for the two models. In particular, an informative prior on n_0 could permit to recover identifiability between the two classes. Unfortunately, in our analysis we do not have any prior information, and non informative priors would lead us to identical weights for the two options.

3.4.1 MCMC for the behavioral effect

One–inflated models (3.2) associated to the behavioral effect can be formalized in terms of right-censored data. In fact, for all units captured once affected by the behavioral effect, we observed a lower bound of the potential number of captures. As such, we have several possibilities of estimation in a Bayesian context. We propose a Gibbs sampler exploiting the latent variables Y^*. The conditional distribution of Y when $Y^* = j$ is concentrated on j when $j \leq 1$, while, for $j > 1$, we have:

$$Y = \begin{cases} 1 & \text{with probability } \omega; \\ j & \text{with probability } 1 - \omega. \end{cases}$$

In our Bayesian approach we evaluate the posterior distribution of the number of units whose latent number of captures Y^* is affected by the behavioral mechanism, (which is a portion of the observed number of units captured once), and their latent number of captures (a passage we will call “imputation” of Y^*). To do that, we consider the conditional distribution of Y^* when $Y = 1$:

$$P(Y^* = j \mid Y = 1, \theta, \omega) = \begin{cases} 0 & \text{if } j = 0; \\ \dfrac{f^*(1|\theta)}{f^*(1|\theta) + \omega(1 - F^*(1|\theta))} & \text{if } j = 1; \\ \dfrac{\omega f^*(j|\theta)}{f^*(1|\theta) + \omega(1 - F^*(1|\theta))} & \text{if } j > 1; \end{cases} \tag{3.5}$$

where F^* denotes the cumulative distribution function associated to f^*. Then, the probability of a unit captured once of being affected by the behavioral effect and contributing to the one–inflation is equal to

$$P(Y^* > 1|Y = 1) = \frac{\omega(1 - F^*(1|\theta))}{f^*(1|\theta) + \omega(1 - F^*(1|\theta))}. \tag{3.6}$$

As a consequence, the distribution of Y^* for those units is truncated in 0 and 1:

$$P(Y^* = j|Y = 1, B = 1) = \frac{f^*(j \mid \theta)}{1 - F^*(1 \mid \theta)} \quad \text{for } j \geq 2. \tag{3.7}$$

Denote as $\hat{n}$ the number of units for which $Y^* > Y$, and as $\hat{n}_j$ the number of such units having j imputed captures, so that $\sum_{j>1} \hat{n}_j = \hat{n}$. Denote as n_j^* the total number of units in the population captured j times after the imputation step, that is,

$$n_j^* = \begin{cases} n_0 & \text{for } j = 0; \\ n_1 - \hat{n} & \text{for } j = 1; \\ n_j + \hat{n}_j & \text{for } j > 1. \end{cases}$$

Denote as n_j^{*i} the analogous number of units in the i–th component of the mixture. Then, the algorithm for a one–inflated DPM of Poissons comprises five steps formally identical to the steps of Section 3.3.1 with f^*, n_j^{*i}, n_j^* and n^{*i} substituting f, n_j^i , n_j and n^i, and the following additional steps:

6. Update the number $\hat{n}$ of units affected by the behavior effect, which, conditional on the current value of θ, by (3.6), can be generated as

$$\hat{n} \sim Binom\left(n_1\,,\ \frac{\omega(1 - F^*(1|\theta))}{f^*(1|\theta) + \omega(1 - F^*(1|\theta))}\right). \tag{3.8}$$

7. Generate a number Y^* of latent captures for each of the $\hat{n}$ units from the truncated distribution (3.7), and update accordingly the values $\{n_j^*\}_{j=2,3,\ldots}$.

8. Since we assume a $Beta(\alpha_\omega, \beta_\omega)$ prior over ω, update ω from

$$Beta\Big(\alpha_\omega + \hat{n}\,,\ \beta_\omega + n_{2+}\Big),$$

where n_{2+} denotes the number of observed units captured more than once, $n_{2+} = \sum_{j>1} n_j$, a quantity which remains fixed throughout the procedure.

3.4.2 MCMC for spurious cases: Trimming

Under model (3.3) we assume that a part of the units captured once are spurious, and the observed value n_1 has to be considered as an upper bound of the actual number of singletons, n_1^*, to be estimated according to the baseline distribution. Bayesian estimation of this one–inflated model is simply obtained by updating in the MCMC a value for ψ and one for n_1^*. We will call this procedure "trimming". If we assume a $Beta(\alpha_\psi, \beta_\psi)$ prior over ψ, the full conditionals assume the following forms:

$$n_1^* \sim Binom\left(n_1\,,\ \frac{(1-\psi)f^*(1|\theta)}{\psi + (1-\psi)f^*(1|\theta)}\right),$$

$$\psi \sim Beta\Big(\alpha_\psi + n_1 - n_1^*\,,\ \beta_\psi + n_{2+} + n_0^* + n_1^*\Big).$$

Then, we follow the steps of Section 3.3.1, where n_1 is to be replaced with n_1^*, in step 4 we update n_0 from $NegBin(n, 1-(1-\psi)f^*(0|\theta))$, and in step 5 f is to be replaced with f^*.

3.4.3 MCMC for spurious cases: Discounting

In the presence of possibly erroneous data, some authors suggest the possibility to conduct the analysis solely on the basis of a part of the records, that is, to simply discard the units which are possibly affected by some errors, and derive the estimate of N solely from the remaining, error-free units. The possibility is hinted in Richardson (2015) in the context of spurious units in MSE, is adopted by Bunge et al. (2014) and Willis (2016) in the context of microbial species abundance estimate, and by Böhning & van der Heijden (2019) in the case of one–inflated repeated count data. In practice, in our case, we should discard the observed number of singletons and re-estimate it, alongside with n_0, on the basis of the values $n_j, j > 1$, non affected by inflation. We will call this process "discounting", in accordance with Bunge et al. (2014). We deem useful to highlight two points on this approach: 1) discounting appears to be useful only when the source of one–inflation are spurious cases; 2) discounting implies a certain loss of information with respect to trimming, but it allows to test the one–inflation hypothesis.

As for the first point, note that, if (3.3) holds, and we discard the observed value n_1, we cannot estimate ψ, as we ignore any potential spurious unit. However, the discounted estimate of n_0 would be an estimate of the unobserved units in the subpopulation for which $S = 0$, i.e., an estimate n_0^* under the baseline distribution of (3.3), which coincides with the number of unobserved units under (3.3). On the converse, discounting would not allow an estimate of n_0 under model (3.2). In fact, if (3.2) holds and we ignore n_1, we cannot estimate ω in (3.2), and the discounted estimate of n_0 would be limited to the subpopulation for which $B = 0$, which is just a portion of n_0 under (3.2).

As for the second point, if we do not question the hypothesis behind (3.3), discounting implies a loss of information with respect to trimming, as the observed value of n_1 is no longer held as an upper bound for the number of units captured once. In the cases we are considering, quite often n_1 represents an important portion of the total amount of observed units, so the estimate of the posterior of n_0 can change considerably with respect to the trimming approach. On the other hand, discounting can be used for assessing robustness of the estimate of N to the one–inflation hypothesis. In fact, with this procedure, we admit the possibility of values of the number of singletons higher than the observed n_1, that is, the possibility of the opposite phenomenon to one–inflation. At the same time, if discounting leads to an estimate of n_0 close to that resulting from trimming, it can be considered as a validation of the associated hypothesis of one–inflation.

In our Bayesian approach the joint estimation of n_0^* and n_1^* does not present particular difficulties. In fact, at each iteration of the MCMC, we simply have to generate a value for the couple (n_0^*, n_1^*) from the following distribution:

$$P((n_0^*, n_1^*)|\theta, n_{2+}) \propto P(N)\frac{N!}{n_{2+}!n_0^*!n_1^*!}\left(1 - f^*(0|\theta) - f^*(1|\theta)\right)^{n_{2+}} f^*(0|\theta)^{n_0^*} f^*(1|\theta)^{n_1^*}. \tag{3.9}$$

That is, if we adopt the improper prior $P(N) \propto 1/N$ over N, we simply have to generate from a Negative Multinomial distribution. Then, the Gibbs sampler can proceed according to steps 1., 2., 3. and 5. of Section 3.3.1 where f is replaced by f^*.

3.5 Sparse finite mixtures

An alternative to the Dirichlet process mixtures is represented by the sparse finite mixtures (SFMs), (see Malsiner-Walli et al. (2016) and Malsiner-Walli et al. (2017)). SFMs specify a prior Dirichlet distribution on the weights of a finite mixture distribution: $(\pi_1, \ldots, \pi_k) \sim Dir(e_1, \ldots, e_k)$. The Dirichlet is assumed to be symmetric: i.e., $e_i = e$, for $i = 1, \ldots, k$, and such that it favors sparse distributions, i.e., the value e is chosen to be smaller than 1. We also choose the finite mixture to be overfitting, i.e., the number of components k is assumed larger than what we can reasonably assume the actual number of components in the population may be. In this way, the actual number of components in the data is not fixed a priori, but rather, as for DPM, it is random by construction and can be estimated using MCMC methods. As shown by Green & Richardson (2001), DPM can be seen as the limiting case of a SFM.

In our context, a favorable property of SFMs consists in their behavior when the number of observations n increases. Let us call "cluster" the set of all units allocated to a certain component. Unlike DPM, a SFM avoids to create new clusters as k increases, even if n goes to infinity. In fact, Müller & Mitra (2013) demonstrated that a DPM tends to increase the number of clusters with n, that is, it is very likely that one big cluster is found, the sizes of further clusters geometrically decay, and many clusters consisting of a single unit are estimated. Also Miller & Harrison (2013) discusses the properties of DPM with respect to the number of components, highlighting the risk of overestimating the number of clusters.

We exploit a result from Frühwirth-Schnatter & Malsiner-Walli (2019), which shows that a careful choice of the prior of the precision parameters ϕ of the DPM and the parameter e of the SFM allows to achieve sparse clustering in both models. That is, we avoid overfitting the number of clusters, which remains stable for increasing value of k. In this way, the clustering performance of DPM and SFM become comparable and provide very similar results. In particular, the values suggested by Frühwirth-Schnatter & Malsiner-Walli (2019) to obtain the same level of sparsity are: $e = 1/(20 \cdot k)$ for the SFM, and $\phi \sim Gamma(1, 20)$ for the precision parameter of the DPM.

The MCMC algorithms for SFM models are identical to the those introduced in the Chapter for the DPMs, where the steps for generating the mixing weights are replaced by sampling $(\pi_1, \ldots, \pi_k)$ from $Dir(n^1 + e, \ldots, n^k + e)$.

3.6 Application to criminal populations

In this Section, we apply our models to estimate the number of people implicated in smuggling, prostitution exploiting, and drug trafficking. In all our models we utilized the improper prior $P(N) \propto 1/N$ over N. All DPM models utilized the sparsity prior suggested in Frühwirth-Schnatter & Malsiner-Walli (2019) for the precision parameter, i.e., $\phi \sim Gamma(1, 20)$, and a number of component equal to 10 (larger values for k

have been tested without observing changes in the results, as we show in the Supplemental material). We set a Gamma prior for all parameters λ_i with a common Gamma hyper prior for the component-specific rate parameter chosen as suggested in Frühwirth-Schnatter & Malsiner-Walli (2019), to provide substantial probability to large values:

$$\begin{aligned} \lambda_i|\beta_\lambda &\sim Gamma(0.1, \beta_\lambda), \qquad \text{for } i = 1, \ldots, k, \\ \beta_\lambda &\sim Gamma(0.5, 5\,\bar{y}), \end{aligned}$$

with $\bar{y}$ being the mean number of captures in the observed data. In addition, for the one–inflated models, we set a uniform ($Beta(1, 1)$) prior over the inflation parameter ω or ψ. In all the applications, the number of iterations of the MCMC algorithm is one million with a thinning of 20 observations. Standard diagnostic tools confirmed the convergence of the algorithm in all cases.

We present separately the results deriving from ignoring one–inflation, and from the two one–inflation hypotheses.

3.6.1 Population estimates when ignoring one–inflation

Table 3.1 summarizes the results for the three datasets when we ignore one–inflation. It shows the posterior modes and credible intervals of N under the DPM model, the Poisson and the Negative Binomial (NB) (for a Bayesian approach to the latter two see 2). For the sake of comparison with non–Bayesian approaches, we also report in Table 3.1 the results from the well-known Chao and Zelterman estimators (see Chao (2014) for a detailed description), and from the maximum likelihood non–parametric Poisson mixture model proposed by Norris & Pollock (1996) and Norris & Pollock (1998) (NPML, in the Table). We utilized the R package SPECIES (Wang (2011)) to estimate the NPML model and its confidence interval for N (via bootstrap) and to derive the confidence intervals of Chao estimator.

Table 3.1. Posterior modes and credible intervals of N for smuggling, prostitution exploitation, and drug trafficking data when ignoring one–inflation

	Smugglers		Prostitution exploiters		Drug traffickers	
	n=3349		n=2740		n=34963	
Model	$\hat{N}$	95% CI	$\hat{N}$	95% CI	$\hat{N}$	95% CI
DPM	12093	(10692 – 13592)	10073	(9049 – 11110)	117678	(112996 – 124051)
Poisson	5583	(5392 – 5774)	7223	(6783 – 7693)	54447	(53927 – 54975)
NB	153466	(110478 – 529338)	89140	(35162 – 188368)	1857809	(1003669 – 2191696)
NPML	12018	(9789 – 13233)	10012	(9345 – 11286)	154392	-
Chao	11387	(10451 – 12447)	9851	(8961 – 10868)	106042	(103441 – 108741)
Zelterman	12052	(10952 – 13152)	10030	(9033 – 11027)	111395	(108471 – 114319)

Even if the Negative Binomial might appear as an optimal choice as it represents a simple two-parameters generalization of the Poisson allowing for heterogeneity in the capture probabilities, it has been proved to be hard to use in capture–recapture. In fact, it suffers from the so–called "boundary problem". That is, when in the observed data the mean number of captures is close to one (which is typically the case in the presence of one–inflation), the model severely overestimate the number of uncaptured units, sometimes by several orders of magnitudes. Looking at the results, it seems safe to assert that we incurred in the boundary problem in all three datasets. On the other hand, DPM models do not have this problem, and provide much greater flexibility to model heterogeneity in the data. In the first two datasets the DPM models produce results very similar to those produced by the NPML procedure. This is not surprising, since we used non-informative priors. In the drug traffickers data, we can see a noticeable difference in the estimates for N that is to be ascribed to the different number of components identified by the two methods. Indeed, the DPM identifies 6 non-empty components (see Table 3.4), while the NPML identifies 5 components. In addition, the bootstrap procedure utilized for the NPML is not able to provide a confidence interval in a reasonable time-span. Surprisingly, the DPM models produce results quite close to those deriving from the Zelterman estimator, confirming and motivating its popularity due to its simplicity and its robustness against mis-specification of the Poisson model.

3.6.2 Population estimates when modeling the behavioral effect

Table 3.2 shows the estimates produced by models that take into account one–inflation caused by the behavioral effect. We considered three Bayesian model of the kind (3.2): the one–inflated DPM model of Section 3.4.1 (labeled as DPM in Table 3.2), the one–inflated Poisson and the one-inflated Negative Binomial models proposed in 2 (OIP and OINB respectively in Table 3.2). For the sake of comparison, we considered the frequentist one–inflated Poisson finite mixture models proposed by Godwin (2019), (Inflmix in Table 3.2 - we name it as the R function provided by the author as supplementary material). Unfortunately, the code provided in Godwin (2019) does not provide a confidence interval estimate, and a bootstrap procedure seems to be too cumbersome.

Table 3.2. Posterior modes and credible intervals of N for smuggling, prostitution exploitation, and drug trafficking data when modeling the behavioral effect

	Smugglers n=3349		Prostitution exploiters n=2740		Drug traffickers n=34963	
Model	$\hat{N}$	95% CI	$\hat{N}$	95% CI	$\hat{N}$	95% CI
DPM	5830	(4959 – 8028)	6613	(5284 – 9870)	111708	(91854 – 120860)
OIP	3570	(3523 – 3614)	3885	(3655 – 4153)	36754	(36638 – 36877)
OINB	40256	(26970 – 71858)	19566	(6174 – 71710)	765406	(320886 – 868786)
Inflmix	7078		10671		-	

As expected, when we take into consideration one–inflation by explicitly modeling the behavioral effect, we obtain lower estimates of N than those obtained under the corresponding baseline distributions. Similarly to the previous analysis, the estimates derived by OIP models are by far the lowest in all three datasets. This suggests the presence of unobserved heterogeneity in the capture probabilities which remains after taking into account one–inflation. The formulation of a one–inflated counterpart of the Negative Binomial should also serve, in the intention of the authors, to mitigate the consequences of the boundary problem. Unfortunately, the problem seems to remain in all three datasets. So, the OINB distribution does not help for taking into account the heterogeneity in these applications, and the DPM again appears as a safer choice.

The estimates of N resulting from our one–inflated DPM are always lower than those obtained by the frequentist model proposed by Godwin (2019). This difference is due to the different number of components identified by the frequentist procedure (by means of the Akaike information criterion), which is always one more than these recognized by our model, in both smugglers and prostitution exploiters data. If we fix in the Inflmix procedure the same number of components we found with the DPM, we obtain quite similar estimates of N. We were not able to obtain an estimate for the drug traffickers data with Inflmix, not even by fixing the number of components. These results in our opinion indicate a certain computational advantage of our approach that avoids estimating a different model for each number of components, and easily produces interval estimates for any quantity of interest.

3.6.3 Population estimates when modeling the spurious cases

We presented two estimation procedures for distributions of type (3.3) which model one–inflation caused by spurious cases. Table 3.3 shows the estimates produced by trimming and discounting algorithms, as detailed in Sections 3.4.2 and 3.4.3. We also show the results from the modified Chao's estimator for one–inflated data (mod.Chao in Table 3.3), proposed in Böhning et al. (2018). This modification of Chao's estimator follows a discounting approach, and provides a lower-bound for the population size estimate in case of one–inflated data under a baseline mixture of Poissons distribution.

Table 3.3. Posterior modes and credible intervals of N for smuggling, prostitution exploitation, and drug trafficking data when modeling spurious cases

	Smugglers n=3349		Prostitution exploiters n=2740		Drug traffickers n=34963	
Model	$\hat{N}$	95% CI	$\hat{N}$	95% CI	$\hat{N}$	95% CI
DPM trimming	4526	(3983 – 6303)	4915	(3775 – 8602)	111087	(78562 – 119188)
DPM discounting	4657	(3980 – 6314)	4675	(3634 – 13771)	94539	(76506 – 118737)
mod.Chao	4431	(3943 – 4919)	6493	(4163 – 8823)	54302	(51601 – 57004)

As expected, when we take into consideration one–inflation caused by spurious cases, we obtain population size estimates which are lower than those obtained under the corresponding distributions modeling the behavioral effect.

It is worthwhile noting that in Table 3.3 the estimates of N are obtained by using the observed n_1, so that the variability depends only on the posterior distribution of n_0. In this way, we have a fair comparison with Tables 3.1 and 3.2 as we can appreciate the differences that the three hypotheses lead to the estimates of n_0. That being said, if one is confident that the source of one–inflation is actually the presence of spurious cases, their count should be detracted by the total population size estimates. That is, we should consider the posterior distribution of the quantity $n_0^* + n_1^* + \sum_{j>1} n_j$. This operation does not present particular difficulties in our approach, since both the trimming and the discounting algorithms provide estimates for the posterior distribution of n_1^*.

As anticipated in Section 3.4.3, the comparison between trimming and discounting results can help us in assessing the hypothesis of spurious cases one–inflation. To this purpose, it is illustrative the comparison between the posterior distribution of the non-spurious singleton counts n_1^* provided by the two algorithms in Figure 3.2. The solid vertical line corresponds to the observed n_1.

For data on smuggling (top panel of Figure 3.2), the posterior distributions resulting from the two algorithms are quite close to each others, as confirmed by the estimates in Table 3.3, and considerably far from the observed value n_1, which is not included in the credible intervals. This result firmly corroborates the assumption of one-inflation. The results for data on prostitution exploitation (central panel of Figure 3.2), are slightly less conclusive: the two distributions have similar modes, but that coming from discounting has more mass on the right tail. As already noted, n_1 is an upper bound for the value n_1^* estimated by trimming, while the estimates from discounting can exceed it, suggesting in this case a small evidence of one–deflation. This is confirmed by the upper bound of the credible interval for N, which is much far to the right than that of the trimming. An opposite result is found on the data on drug trafficking, on the bottom of Figure 3.2. The posterior

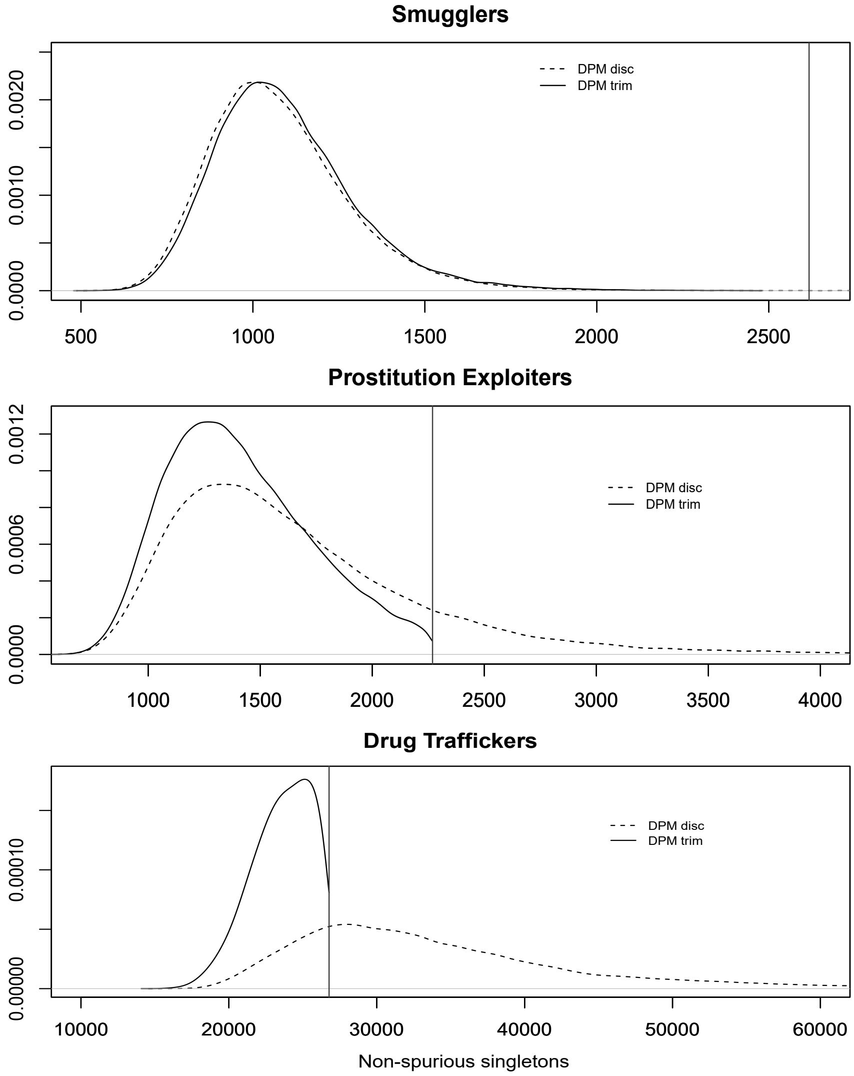


Figure 3.2. Posterior distributions of non–spurious singleton counts n_1^* under DPM for spurious cases with trimming (solid line), and discounting (dashed line). Vertical lines indicate the observed value n_1.

mode of n_1^* under trimming is close to n_1, suggesting a small amount of one–inflation, which is confirmed by the value $\hat{\psi}$ in Table 3.4. However, the posterior distribution of n_1^* under discounting presents much larger variability with a rather long right tail, and seems to favour a one–deflation hypothesis. This result seems to indicate that the small amount of one–inflation found with trimming could be non significant and, more generally, indicate a larger uncertainty in the model results. In Figure 3.3 we show the posterior distributions of N under the DPM model without one inflation (solid line), one–inflated DPM modeling behavioral effect (dotted line), and one–inflated DPM modeling spurious cases (dashed line), using trimming algorithm. The regular shape of the posterior distributions in case of non inflated models is evident from Figure 3.3 for all the data sets, so the differences in adopting the posterior mode, median or mean are quite negligible. Regularity of the posterior distributions is also observed for the one–inflated models in the case of smuggling data, top panel in Figure 3.3, while it does not hold for drug trafficking data, bottom panel in Figure

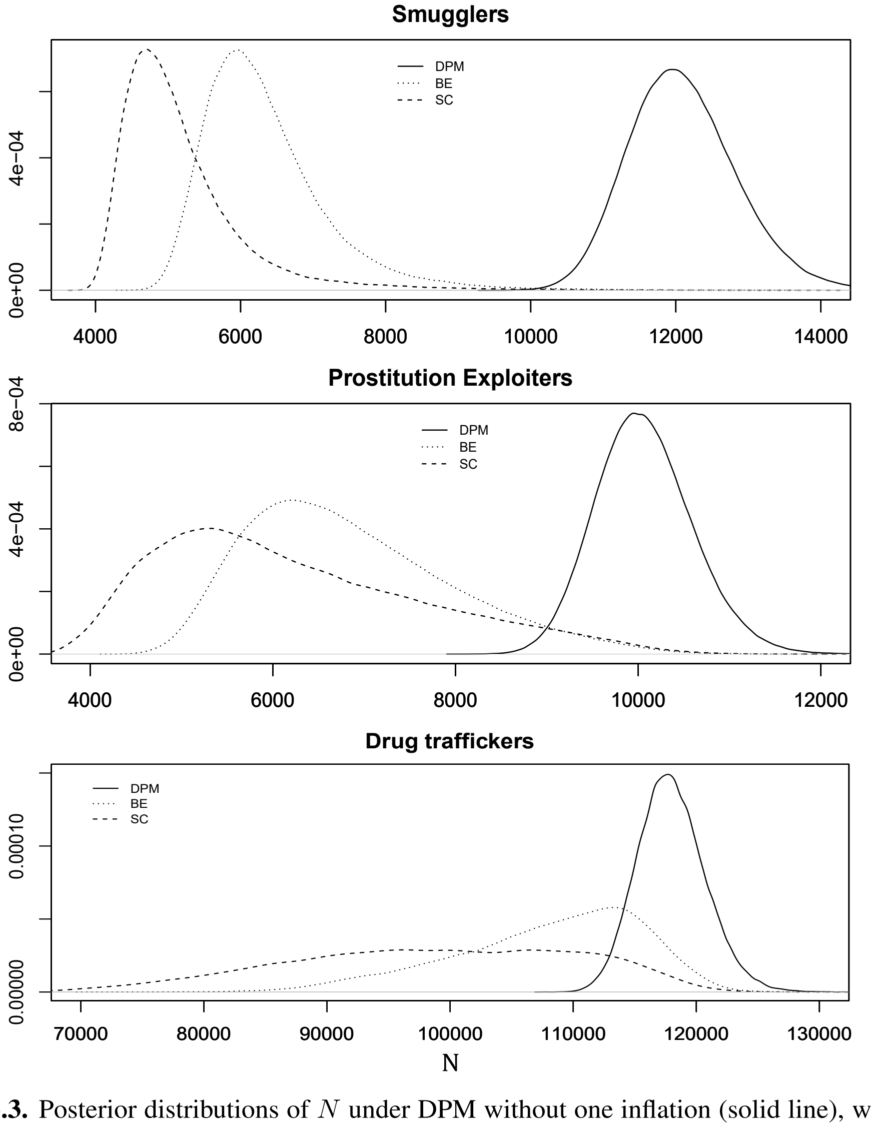


Figure 3.3. Posterior distributions of N under DPM without one inflation (solid line), with one–inflation due to behavioral effect (dotted line), and by spurious cases (dashed line)

Table 3.4. Posterior means of λs and πs for each non empty component of DPM models

	Parameter	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6
			Smugglers				
w/o one–inflation	λ_i	0.297	3.523	13.055			
	π_i	0.970	0.026	0.003			
behavioral effect	λ_i	0.715	4.657	14.212			
$\hat{\omega}$=0.43	π_i	0.937	0.054	0.009			
spurious cases	λ_i	0.709	4.702	14.378			
$\hat{\psi}$=0.28	π_i	0.938	0.053	0.009			
			Prostitution Exploiters				
w/o one–inflation	λ_i	0.310	3.945				
	π_i	0.991	0.009				
behavioral Effect	λ_i	0.471	4.574				
$\hat{\omega}$=0.26	π_i	0.988	0.012				
spurious cases	λ_i	0.508	4.727				
$\hat{\psi}$=0.16	π_i	0.987	0.013				
			Drug traffickers				
w/o one–inflation	λ_i	0.316	2.648	7.060	14.950	29.113	49.742
	π_i	0.962	0.030	0.006	0.002	0.001	0.0002
behavioral Effect	λ_i	0.353	2.750	7.147	15.020	29.309	49.888
$\hat{\omega}$= 0.06	π_i	0.958	0.033	0.006	0.002	0.001	0.0002
spurious cases	λ_i	0.360	2.777	7.229	15.092	29.250	49.742
$\hat{\psi}$=0.03	π_i	0.957	0.034	0.006	0.002	0.001	0.0002

3.3. As expected, if we ignore one–inflation, we risk severely overestimating the population size and one–inflation due to spurious cases produces lower estimates than one–inflation due to the behavioral effect. Unfortunately, due to the lack of additional information, we are not able to conclude a definite answer about the population size estimates, however, we believe that the proposed models and the performed analysis shed some lights on the need of considering a wider range of estimated values in the presence of one–inflation.

3.6.4 Number of components and sparsity

As introduced in Section 3.5, the use of a sparsity prior for the precision parameter ϕ allows us to avoid overfitting the number of clusters for DPM. In addition, in this way we ensure that DPM models produce fully comparable results with SFM models. We firstly verified that different values of k do not affect the number of components we detect, and actually with prior on ϕ favoring sparse mixtures we always identified the same number of components and the parameters estimates are not affected by the choice of k. These results allow us to safely choose a relatively small value for the truncated number of components k, with the two-fold advantage of reducing the computational complexity of the DPM algorithm and of avoiding model selection with respect to the number of components as in traditional mixture models. In Figure 3.4 we show, by way of example for the drug trafficking data, box–plots of posterior estimates $\hat{N}$ by different choices of k in the DPM, with and without considering one–inflation. The Figure clearly shows that the population size estimates become stable once we select $k \geq 6$, since the number of non-empty components is consistently estimated at six, even when k increases. Similar graphics are observed with the smuggling and the prostitution exploitation data, where the number of non-empty components are estimated at three and two, respectively. We also

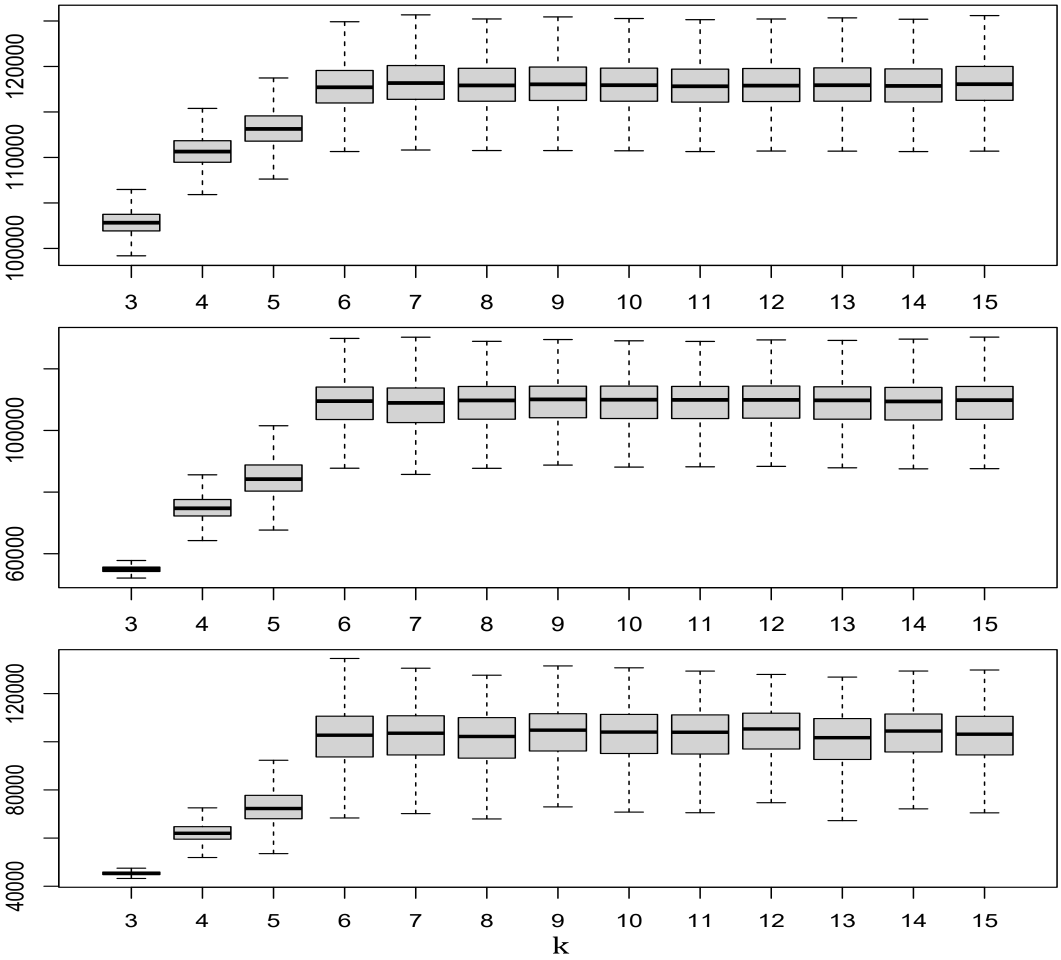


Figure 3.4. Box–plots of the generated posterior distributions of N under DPM by different values of k for drug traffickers data. Top, without one–inflation; middle, modeling the behavioral effect; bottom, modeling spurious cases

verified that the components detected by the DPM and the corresponding resulting estimates are the same as obtained with the SFM, and, as a matter of fact, the posteriors of N are almost identical in all cases.

In Figure 3.5 we show, by way of example for the drug trafficking data, the comparison between the posterior distributions for the population size produced by DPM and SFM, with and without considering one–inflation. The Figure confirms the equivalence between the two models, once the precision parameter of the DPM is chosen appropriately. Similar results are obtained with the smugglers and the prostitution exploiters data.

3.7 Concluding remarks

In this work we have proposed a Bayesian capture–recapture approach to estimate the number of individuals involved in three types of criminal activities in Italy. We have identified three possible sources of one–inflation for the data at our disposal, namely, a behavioral effect, the presence of spurious cases, and linkage errors in recognizing the units, and we have modeled, compared and discussed the first two. To handle population

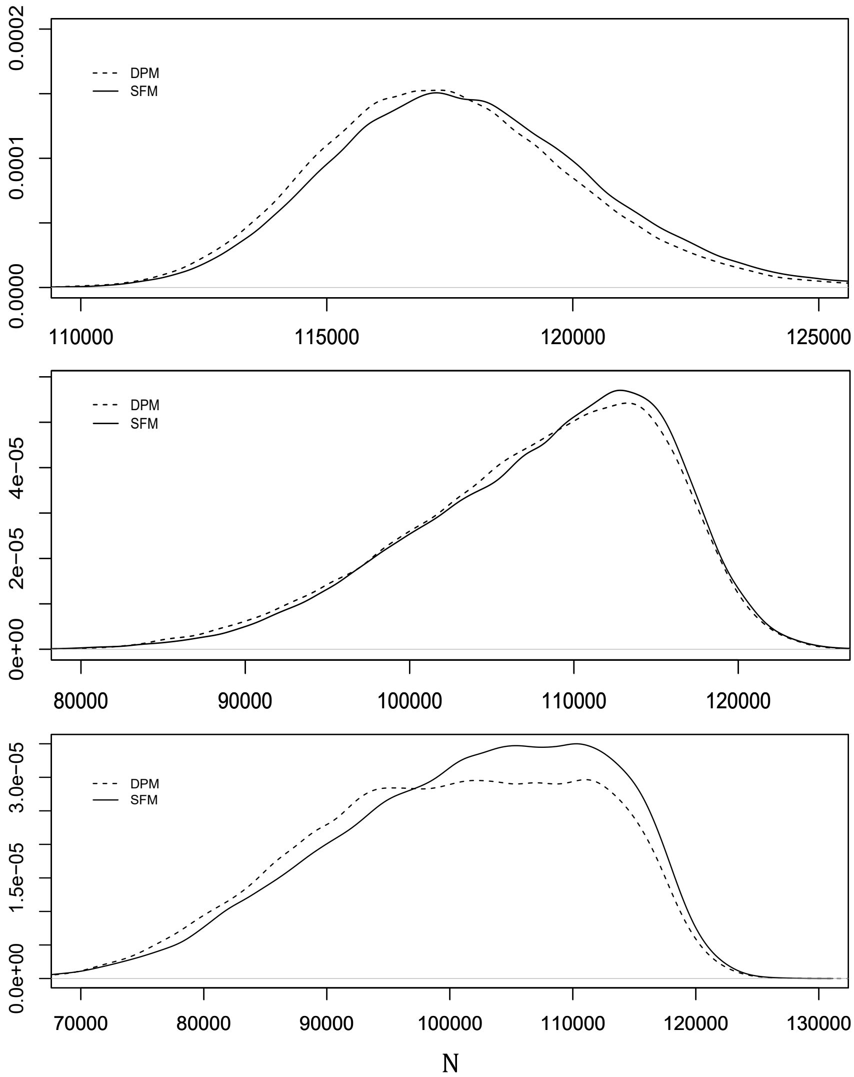


Figure 3.5. Estimated posterior distributions of N under DPM and SFM for drug traffickers data. Top- without one–inflation; middle- modeling the behavioral effect; bottom- modeling spurious cases.

heterogeneity, we have considered a semi-parametric approach based on Dirichlet process mixture to automatically take into account the uncertainty over the number of components necessary to fit the data. The resulting models consist in one-inflated semi-parametric mixtures, which are new in the capture–recapture literature.

We suspect that our data may be affected by both one–inflation mechanisms, and we have seen that it is not possible to distinguish between the two on the basis of the likelihood of the two models. Our Bayesian approach allows the inclusion of prior information that could direct us towards one of the two options, but, unfortunately, in our applications we had to resort to uninformative priors. As a consequence, we have to compare the results of both one–inflated models and consider a larger range of possibilities for the population size estimates. It is worth mentioning that there are capture-recapture applications for which the origin of the one–inflation can be certainly recognized, e.g., in species abundance for microbial ecology, where a behavioral effect is excluded, and our models for spurious cases can certainly be employed.

Even if we do not have any prior information on the population at hand, our Bayesian approach seems to have some computational advantage over non–Bayesian alternatives. We saw in Table 3.1 and 3.2 that the MCMC allows us to reliably estimate up to 6 components providing at the same time the interval estimates for all the parameters.

We are currently working on extensions of these models to cope with the linkage and deduplications errors. A Bayesian approach for record linkage and deduplication problems is provided by Steorts et al. (2016). In our context, missing links, i.e., false non matches, induce one-inflation reducing the frequency of multiple captures, while false matches operate in the opposite direction inducing one-deflation and a larger frequency of multiple captures. Note that both errors introduce additional uncertainty on the total number of distinct units in the sample, which, as a consequence, should be considered as a random quantity. Ideally, it would be important to have access to the raw data. In this way, we could take into account the whole record linkage process uncertainty in population size estimation, for instance via a hierarchical structure, as in Tancredi & Liseo (2011), Sadinle (2018), and Tancredi et al. (2020), in order to propagate the uncertainty between the parameter estimation step and the matching procedure.

References

Böhning, D. (2015), 'Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling', *Metron* **73**(2), 201–216.

Böhning, D., Dietz, E., Kuhnert, R. & Schön, D. (2005), 'Mixture models for capture-recapture count data', *Statistical Methods and Applications* **14**(1), 29–43.

Böhning, D. & Friedl, H. (2021), 'Population size estimation based upon zero-truncated, one-inflated and sparse count data', *Statistical Methods & Applications* pp. 1–21.

Böhning, D., Kaskasamkul, P. & van der Heijden, P. (2018), 'A modification of Chao's lower bound estimator in the case of one-inflation', *Metrika* **82**(3), 361–384.

Böhning, D. & Ogden, H. E. (2021), 'General flation models for count data', *Metrika* **84**(2), 245–261.

Böhning, D. & Schön, D. (2005), 'Nonparametric maximum likelihood estimation of population size based on the counting distribution', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(4), 721–737.

Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W. & Viwatwongkasem, C. (2004), 'Estimating the number of drug users in Bangkok 2001: A capture–recapture approach using repeated entries in one list', *European journal of epidemiology* **19**(12), 1075.

Böhning, D. & van der Heijden, P. G. (2019), 'The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in britain', *The Annals of Applied Statistics* **13**(2), 1198–1211.

Borchers, D., Buckland, S., Stephens, W. & Zucchini, W. (2002), *Estimating animal abundance: closed populations*, Vol. 13, Springer Science & Business Media.

Bunge, J., Willis, A. & Walsh, F. (2014), 'Estimating the number of species in microbial diversity studies', *Annual Review of Statistics and Its Application* **1**, 427–445.

Chambers, R. (2009), 'Regression analysis of probability-linked data', *Official Statistics Research Series –Statistics New Zealand* **4**.

Chambers, R. C. & Kim, G. (2015), Secondary analysis of linked data, *in* K. H. H. Goldstein & C. Dibben), eds, 'Methodological Developments in Data Linkage', p. Chapter 5.

Chambers, R. L. & da Silva, A. D. (2019), *Improved secondary analysis of linked data: a framework and an illustration*, Series A, Journal of the Royal Statistical Society.

Chao, A. (2014), 'Capture-recapture for human populations', *Wiley StatsRef: Statistics Reference Online* pp. 1–16.

Chib, S. (1995), ‘Marginal likelihood from the Gibbs output’, *Journal of the American Statistical Association* **90**(432), 1313–1321.

Chib, S. & Jeliazkov, I. (2001), ‘Marginal likelihood from the metropolis–hastings output’, *Journal of the American Statistical Association* **96**(453), 270–281.

Chipperfield, J. O., Bishop, G. R. & Campell, P. (2011), ‘Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data’, *Survey Methodology* **37**, 13–24.

Chipperfield, J. O. & Chambers, R. C. (2015), ‘Using bootstrap to account for linkage errors when analysing probabilistically linked categorical data’, *Journal of Official Statistics* **31**, 397–414.

Chiu, C.-H. & Chao, A. (2016), ‘Estimating and comparing microbial diversity in the presence of sequencing errors’, *PeerJ* **4**, e1634.

Christen, P. (2012), ‘A survey of indexing techniques for scalable record linkage and deduplication’, *ISEE Transactions on Knowledge and Data Engineering* **24**.

Copas, J. B. & Hilton, F. J. (1990), ‘Record linkage: Statistical models for matching computer records’, *Journal of the Royal Statistical Society, Ser. A* **153**, 287–320.

Creel, S., Spong, G., Sands, J. L., Rotella, J., Zeigle, J., Joe, L. & Smith, D. (2003), ‘Population size estimation in yellowstone wolves with erro-prone noninvasive microsatellite genotypes’, *Molecular ecology* **12**(7), 2003–2009.

Cruyff, M. J. & van der Heijden, P. G. (2008), ‘Point and interval estimation of the population size using a zero-truncated negative binomial regression model’, *Biometrical Journal* **50**, 1035–1050.

Di Cecco, D. (2019), Estimating population size in multiple record systems with uncertainty of state identification, *in* ‘Analysis of Integrated Data’, Chapman and Hall/CRC, pp. 169–196.

Di Cecco, D., Di Zio, M., Filipponi, D. & Rocchetti, I. (2018), ‘Population size estimation using multiple incomplete lists with overcoverage’, *Journal of Official Statistics* **34**(2), 557–572.

Di Consiglio, L. & Tuoto, T. (2018), ‘Population size estimation and linkage errors: the multiple lists case’, *Journal of official statistics* **34**(4), 889–908.

Di Consiglio, L., Tuoto, T. & Zhang, L.-C. (2019), Capture-recapture methods in the presence of linkage errors, *in* ‘Analysis of Integrated Data’, Chapman and Hall/CRC, pp. 39–71.

Enamorado, T., Fifield, B. & Imai, K. (2019), ‘Using a probabilistic model to assist merging of large-scale administrative records’, *American Political Science Review* **113**, 353–371.

Farcomeni, A. (2020), ‘Population size estimation with interval censored counts and external information: Prevalence of multiple sclerosis in rome’, *Biometrical Journal* **62**(4), 945–956.

Farcomeni, A. & Scacciatelli, D. (2013), ‘Heterogeneity and behavioral response in continuous time capture–recapture, with application to street cannabis use in italy’, *The Annals of Applied Statistics* **7**(4), 2293–2314.

Fegatelli, D., Farcomeni, A. & Tardella, L. (2017), Bayesian population size estimation with censored counts, *in* 'Capture-recapture methods for the social and medical sciences', Chapman and Hall/CRC, pp. 371–385.

Fegatelli, D. & Tardella, L. (2018), 'Moment-based Bayesian Poisson Mixtures for inferring unobserved units', *arXiv preprint arXiv:1806.06489* .

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

Frühwirth-Schnatter, S. & Malsiner-Walli, G. (2019), 'From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering', *Advances in data analysis and classification* **13**(1), 33–64.

Godwin, R. (2017), 'One-inflation and unobserved heterogeneity in population size estimation', *Biometrical Journal* **59**(1), 79–93.

Godwin, R. (2019), 'The one-inflated positive Poisson mixture model for use in population size estimation', *Biometrical Journal* **61**(6), 1541–1556.

Godwin, R. & Böhning, D. (2017), 'Estimation of the population size by using the one-inflated positive Poisson model', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66**(2), 425–448.

Goldstein, H., Harron, K. & Wade, A. (2012), 'The analysis of record-linked data using multiple imputation with data value priors', *Statistics in Medicine* **31**, 3481–3493.

Green, P. (1995), 'Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–732.

Green, P. J. & Richardson, S. (2001), 'Modelling heterogeneity with and without the Dirichlet process', *Scandinavian Journal of Statistics* **28**(2), 355–375.

Guindani, M., Sepúlveda, N., Paulino, C. & Müller, P. (2014), 'A Bayesian semi-parametric approach for the differential analysis of sequence counts data', *Journal of the Royal Statistical Society. Series C, Applied Statistics* **63**(3), 385.

Gutman, R., Afendulis, C. C. & Zaslavsky, A. M. (2013), 'A Bayesian procedure for file linking to analyze end-of-life medical costs', *Journal of the American Statistical Association* **108**, 34–47.

Gutman, R., Sammartino, C. J., Green, T. C. & Montague, B. T. (2015), 'Error adjustments for file linking methods using encrypted unique client identifier (euci) with application to recently released prisoners who are hiv+', *Statistics in Medicine* **35**, 115–129.

Han, Y. & Lahiri, P. (2018), *Statistical analysis with linked data*, International Statistical Review.

Harron, K., Gilbert, K., Cromwell, D. & van der Meulen, J. (2016), 'Linking data for mothers and babies in de-identified electronic health data', *PLoS ONE* **11**, 10.

Harron, K., Goldstein, H. & Dibben, C. (2015), *Methodological Developments in Data Linkage*, Wiley.

Hausman, J. A. (1978), 'Specification tests in econometrics', *Econometrica* **46**, 1251–1271.

Heasman, M. & (2011), F. (n.d.), *Essnet D. I. – Simulated data for the on the job training*. **URL:** *https://ec.europa.eu/eurostat/cros/content/job-training$_e n$*

Herzog, T. N., Scheuren, F. J. & Winkler, W. E. (2007), *Data Quality and Record Linkage Techniques*, Springer.

Hof, M. H. P. & Zwinderman, A. H. (2012), 'Methods for analysing data from probabilistic linkage strategies based on partially identifying variables', *Statistics in Medicine* **31**, 4231–4242.

Ishwaran, H. & James, L. F. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**(453), 161–173.

Jaro, M. A. (1989), 'Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida', *Journal of the American Statistical Association* **84**, 414–420.

Kass, R. & Raftery, A. (1995), 'Bayes factor and model uncertainty', *Journal of the American Statistical Association* **90**(430), 773–795.

Kim, G. & Chambers, R. C. (2012), 'Regression analysis under incomplete linkage', *Comutational Statistics and Data Analysis* **56**, 2756–2770.

Lahiri, P. & Larsen, M. D. (2005), 'Regression analysis with linked data', *Journal of the American Statistical Association* **100**, 222–230.

Link, W. A. (2003), 'Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities', *Biometrics* **59**(4), 1123–1130.

Link, W. A. (2006), 'Rejoinder to "On Identifiability in Capture-Recapture Models"', *Biometrics* **62**(3), 936–939.

Link, W. A., Yoshizaki, J., Bailey, L. L. & Pollock, K. H. (2010), 'Uncovering a latent multinomial: analysis of mark–recapture data with misidentification', *Biometrics* **66**(1), 178–185.

Malsiner-Walli, G., Frühwirth-Schnatter, S. & Grün, B. (2016), 'Model-based clustering based on sparse finite Gaussian mixtures', *Statistics and computing* **26**(1-2), 303–324.

Malsiner-Walli, G., Frühwirth-Schnatter, S. & Grün, B. (2017), 'Identifying mixtures of mixtures using Bayesian estimation', *Journal of Computational and Graphical Statistics* **26**(2), 285–295.

Manrique-Vallier, D. (2016), 'Bayesian population size estimation using Dirichlet process mixtures', *Biometrics* **72**(4), 1246–1254.

Marchant, N. G., Steorts, R. C., Kaplan, A. & Rubinstein, B. I. P., E. (2019), *D*, Distributed End-to-End Bayesian Entity Resolution, N. d-blink.
URL: *https://arxiv.org/pdf/1909.06039.pdf*

McClintock, B. T., Bailey, L. L., Dreher, B. P. & Link, W. A. (2014), 'Probit models for capture–recapture data subject to imperfect detection, individual heterogeneity and misidentification', *The Annals of Applied Statistics* **8**(4), 2461–2484.

McCrea, R. & Morgan, B. (2014), *Analysis of capture-recapture data*, CRC Press.

Miller, J. W. & Harrison, M. T. (2013), 'A simple example of Dirichlet process mixture inconsistency for the number of components', *Advances in neural information processing systems* p. 199–206.

Müller, P. & Mitra, R. (2013), 'Bayesian nonparametric inference–why and how', *Bayesian Analysis (Online)* **8**(2).

Neter, J., Maynes, E. S. & Ramanathan, R. (1965), 'The effect of mismatching on the measurement of response error', *Journal of the American Statistical Association* **60**, 1005–1027.

Nolan, J. J., Haas, S. M. & Napier, J. S. (2011), 'Estimating the impact of classification error on the "statistical accuracy" of uniform crime reports', *Journal of Quantitative Criminology* **27**(4), 497–519.

Norris, J. L. & Pollock, K. H. (1996), 'Nonparametric mle under two closed capture-recapture models with heterogeneity', *Biometrics* pp. 639–649.

Norris, J. L. & Pollock, K. H. (1998), 'Non-parametric mle for poisson species abundance models allowing for heterogeneity between species', *Environmental and Ecological Statistics* **5**(4), 391–402.

Overstall, A. M., King, R., Bird, S. M., Hutchinson, S. J. & Hay, G. (2014), 'Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in scotland', *Statistics in medicine* **33**(9), 1564–1579.

Owen, A., Jones, P. & Ralphs, M. (2015), Large-scale linkage for total populations in official statistics, *in* 'Methodological Developments in Data Linkage (eds', K. Harron, H. Goldstein and C. Dibben), Chapter 8.

Pledger, S., Pollock, K. H. & Norris, J. L. (2003), 'Open capture-recapture models with heterogeneity: I. cormack-jolly-seber model', *Biometrics* **59**(4), 786–794.

Quenouille, M. (1949), 'A relation between the logarithmic, Poisson, and negative binomial series', *Biometrics* **5**(2), 162–164.

Richardson, C. (2015), 'A note on modelling incomplete contingency tables with censored cells', *Statistics in medicine* **34**(3), 539–540.

Rissanen, J. (1983), 'A universal prior for integers and estimation by minimum description length', *The Annals of statistics* **11**(2), 416–431.

Roberts Jr, J. & Brewer, D. (2006), 'Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture–recapture method', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(4), 745–756.

Rosman, D. L. (2001), 'The western australian road injury database (1987–1996): ten years of linked police, hospital and death records of road crashes and injuries', *Accident Analysis Prevention* **33**(1), 81–88.

Rossmo, D. & Routledge, R. (1990), 'Estimating the size of criminal populations', *Journal of quantitative criminology* **6**(3), 293–314.

Sadinle, M. (2014), 'Detecting duplicates in a homicide registry using a Bayesian partitioning approach', *Annals of Applied Statistics* **8**, 2404–2434.

Sadinle, M. (2017), 'Bayesian estimation of bipartite matchings for record linkage', *Journal of the American Statistical Association* **112**, 600–612.

Sadinle, M. (2018), 'Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations', *The Annals of Applied Statistics* **12**(2), 1013–1038.

Scheuren, F. & Winkler, W. E. (1993), 'Regression analysis of data files that are computer matched', *Survey Methodology* **19**, 39–58.

Scheuren, F. & Winkler, W. E. (1997), 'Regression analysis of data files that are computer matched – part ii', *Survey Methodology* **23**, 157–165.

Seybolt, T. B., Aronson, J. D. & Fischhoff, B. E. (2013), *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict*, Oxford University Press.

Steorts, R. C., Hall, R. & Fienberg, S. E. (2016), 'A bayesian approach to graphical record linkage and deduplication', *Journal of the American Statistical Association* **111**(516), 1660–1672.

Steorts, R., Hall, R. & Fienberg, S. (2017), 'A Bayesian approach to graphical record linkage and de-duplication', *Journal of the American Statistical Association* **111**, 1660–1672.

Tajuddin, R. R. M., Ismail, N. & Ibrahim, K. (2021), 'Estimating population size of criminals: A new Horvitz–Thompson estimator under One-Inflated Positive Poisson–Lindley model', *Crime & Delinquency* p. 00111287211014158.

Tancredi, A. & Liseo, B. (2011), 'A hierarchical Bayesian approach to record linkage and population size problems', *The Annals of Applied Statistics* **5**, 1553–1585.

Tancredi, A., Steorts, R. & Liseo, B. (2020), 'A unified framework for de-duplication and population size estimation (with discussion)', *Bayesian Analysis* **15**(2), 633–682.

Tardella, L. (2002), 'A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity', *Biometrika* **89**(4), 807–817.

Tuoto, T. (2016), 'New proposal for linkage error estimation', *Statistical Journal of the IAOS* **32**(2), 1–8.

UNODC (2015), 'International classification of crime for statistical purposes, version 1.0'.

van der Heijden, P. G., Cruyff, M., & D., B. (2014), 'Capture recapture to estimate criminal populations', *Encyclopedia of criminology and criminal justice* pp. 267–276.

Viwatwongkasem, C., Kuhnert, R. & Satitvipawee, P. (2008), 'A comparison of population size estimators under the truncated count model with and without allowance for contaminations', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50**(6), 1006–1021.

Wang, J.-P. (2011), 'SPECIES: an R package for species richness estimation', *Journal of Statistical Software* **40**(9), 1–15.

Wang, X., He, C. Z. & Sun, D. (2007), 'Bayesian population estimation for small sample capture-recapture data using noninformative priors', *Journal of Statistical Planning and Inference* **137**(4), 1099–1118.

Willis, A. (2016), ‘Species richness estimation with high diversity but spurious singletons’, *arXiv preprint arXiv:1604.02598* .

Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E. & Gleeson, D. M. (2009), ‘Incorporating genotype uncertainty into mark–recapture-type models for estimating abundance using dna samples’, *Biometrics* **65**(3), 833–840.

Xu, C., Sun, D. & He, C. (2014), ‘Objective Bayesian analysis for a capture–recapture model’, *Annals of the Institute of Statistical Mathematics* **66**(2), 245–278.

Zhang, G. & Campbell, P. (2012), ‘Data survey: Developing the statistical longitudinal census dataset and identifying its potential uses’, *Australian Economic Review* **45**, 125–133.

Zhang, L.-C. (2019), On secondary analysis of datasets that cannot be linked without errors, *in* L.-C. Zhang & R. L. Chambers, eds, ‘editors, Analysis of Integrated Data’, Chapman and Hall, London, CRC.

Zhou, M. & Carin, L. (2015), ‘Negative binomial process count and mixture modeling’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(2), 307–320.