A system of population estimates compiled from administrative data only

John Dunne †
Central Statistics Office (CSO), Cork, Ireland.
E-mail: John.Dunne@cso.ie
Li-Chun Zhang
University of Southampton, Southampton, United Kingdom.

Summary. This paper presents a novel system of annual Population Estimates Compiled from Administrative Data Only (PECADO) for Ireland in the absence of a Central Population Register. The system is entirely based on data originated from administrative sources, so that population estimates can be produced even without purposely designed coverage surveys or a periodic census to recalibrate estimates. It requires several extensions to the traditional Dual System Estimation (DSE) methodology, including a restatement of the underlying assumptions, a trimmed DSE method for dealing with erroneous enumerations in the administrative register, and a test for heterogeneous capture probabilities to facilitate the choice of blocking in applications. The PECADO estimates for years 2011 - 2016 are compared to the Census counts in 2011 and 2016. We demonstrate how the system can be used to investigate the Census 2016 undercount in Ireland, in place of the traditional approach of deploying additional population coverage surveys.

Keywords: Signs-of-Life approach, capture-recapture methods, census transformation, erroneous records, under-coverage, heterogeneous capture rates

1. Introduction

1.1. Background - Why a new system of population estimates?

For countries that do not have a high-quality Central Population Register from which demographic statistics can be compiled, such as the case in Scandinavia and a handful of other European countries, the production of reliable demographic statistics on population counts and migration flows can prove challenging. This is particularly true for those countries that have relatively highly variable migration flows that are difficult to estimate. Ireland is one such country.

The Central Statistics Office, Ireland (CSO) enumerated 4.76 million people living in the Republic of Ireland in 2016, at an associated cost in excess of €60m or over €12 per person. The approach to population estimates in the following years is the demographic (or cohort) component method, including a recalibration of the intercensal estimates after the next census. In a Eurostat review of 31 countries published in 2003

†Views and opinions expressed are those of the authors and not necessarily those of CSO, Ireland.

Table 1. Population estimates and their components for Ireland (thousands).
Source: Central Statistics Office, Ireland (http://www.cso.ie).

Year	2011 to 2012	2012 to 2013	2013 to 2014	2014 to 2015	2015 to 2016
Population at time point 1	4,574.9	4,593.7	4,614.7	4,645.4	4,687.8
plus Births	73.2	69.4	68.4	66.4	65.4
minus Deaths	28.7	29.8	29.2	29.9	29.8
plus Immigrants	57.3	62.7	66.5	75.9	82.3
minus Emigrants	83.0	81.3	75.0	70.0	66.2
Population at time point 2	4,593.7	4,614.7	4,645.4	4,687.8	4,739.6

(EUROSTAT, 2003), 19 countries were identified as using the component method for population estimates. A subsequent Eurostat review in 2015 (EUROSTAT, 2015) found that 31 of 44 countries depended on the Census for annual population estimates and of these 31 countries only 9 supplemented their population estimates with information from registers.

This demographic component method can be summarised as follows. To estimate the population at timepoint 2, start with the population estimate at timepoint 1, subtract the estimated deaths and persons emigrated and add the estimated births and persons immigrated in the period between timepoints 1 and 2, and then by ageing the population forward from timepoint 1 to timepoint 2, an estimate of the population is obtained for timepoint 2 in age by sex groups. Population estimates for timepoint 3 are obtained by iterating forward from timepoint 2 in the same manner. The weakness with this approach is that any errors or bias in estimating the components of population change (births, deaths, immigration, emigration) will be carried forward from timepoint to timepoint. These concerns, amplified in the presence of high migration flows, are one of the reasons why some countries such as Ireland undertake a Census at 5 yearly intervals. Table 1 provides an overview of the estimated population and change components for Ireland over the years 2011 to 2016. These estimates are intercensal estimates compiled retrospectively after Census 2016 was completed. In the absence of other information, the post-censal estimates are typically adjusted on a pro rata basis to obtain intercensal estimates. The 2016 Census resulted in a revision to the previous population estimate from 4.67 million to 4.74 million, a net difference of 66,000 persons. The net difference does not reveal the total discrepancy as the component method under-estimates for some cohorts and over-estimates for other cohorts.

This traditional approach to population estimates relies on a costly decennial or quinquennial census in order to recalibrate or benchmark population estimates. The revision of intercensal population estimates relies typically on fairly crude interpolations, which is nevertheless necessary due to the needs of statistical users for these estimates. Every Census instance kicks off a huge costly logistical operation and can cause disruption to other business as usual activities of the statistical agency. The approach has its origins in a time where administrative data systems were not well developed nor generally considered as suitable input to statistical systems. A system of annual population estimates based on administrative data sources will negate the need for these costly and complex operations. Such a system eliminates the need for periodic revisions after each Census

along with the disruption for users and producers of other statistical systems/products that rely on populations estimates. A regular survey may be required to provide reassurance or audit checks for the population estimates based on administrative data or for collecting/validating attribute information (including geography). However such a survey could be conducted with lower cost and far less disruption to existing statistical processes. Provided the survey sample can be linked to the administrative sources, it could also provide useful unit-level information for fine-tuning the population estimation system.

We present a novel system of annual Population Estimates Compiled from Administrative Data Only (PECADO) for Ireland in the absence of a Central Population Register, which does not depend on a periodic census to recalibrate estimates. To the best knowledge of the authors no other country has yet compiled official population estimates solely from administrative data sources where no Central Population Register exists. The development of such a system requires a viable strategy to the creation of two enumeration lists (as the basis of estimation) prepared from administrative data, as well as several extensions to the traditional DSE methodology, including

- a restatement of the underlying assumptions in scenarios where census and coverage surveys are replaced by administrative or other relevant data,
- a trimmed DSE method to deal with erroneous records or overcoverage,
- a test for the homogeneous capture assumption that is required of at least one of the two lists.

The paper then uses this toolkit to demonstrate that the proposed system of population estimates is robust. In doing so, we provide also evidence of undercount in the 2016 Census in Ireland and that the undercount-adjusted Census results are coherent with the PECADO population estimates, once migration estimates are considered to explain difference in the underlying population concepts.

If it is possible to compile reliable population estimates on an annual basis then this would negate the requirement of conducting a periodic census. Ireland could at least move to a decennial Census in line with many other countries providing significant savings to the state. If such a system can be further developed it may make the requirement for a traditional Census obsolete.

The ability to compile reliable population estimates from administrative data sources is a first milestone on any roadmap from a traditional Census to a modern Census based primarily on registers and administrative data. A modern Census holds the promise of being conducted on an annual basis at a fraction of the cost of a traditional Census, as in those countries that have already implemented the so-called register-based census. The implication is relevant to many countries internationally.

The rest of the paper is organised as follows. In Section 2, we outline the fundamental ideas of the PECADO system. Section 3 presents the DSE methodology including extensions as the PECADO toolkit. Section 4 discusses the PECADO estimates in comparison to Census 2011 and 2016, before reconciling the population estimates with official Census 2016 counts in Section 4.3. Section 5 summarises our conclusions and directions of further development of the PECADO system for Ireland.

2. PECADO - The Simple Idea

2.1. If You Don't Have a Central Population Register, Build a Statistical Population Dataset (SPD)

Many countries do not have a Central Population Register. Some of these countries are now actively investigating how to get the benefits of a statistical system based on registers. In the absence of a Central Population Register, the simple idea is to compile a so-called Statistical Population Dataset (SPD) using available data sources as has been explored in UK (ONS UK, 2017) and New Zealand (Dunne and Graham, 2019).

The ideal SPD will have a record for each statistical unit (person) in the target population - each unit identified with a unique identification number. The target population for population estimates requires a person to be living in the State. There will be variations of the basic definition, de facto, de jure, registered etc. but the basic premise is the person is usually resident in the State. Lanzieri (2013) discusses different population concepts in more detail. In compiling an SPD from multiple data sources, 4 main types of error may cause discrepancy to the target population.

- Overcoverage: The SPD has units that do not belong to the target population.
- Undercoverage: The SPD is missing units that belong to the target population.
- Linkage error: Information about different population units in multiple sources are combined into a single record in the SPD.
- Domain misclassification: A domain attribute is incorrect despite correct linkage, such as when a wrong choice is made given conflicting values in different sources.

The CSO, like other statistical agencies, has also compiled an SPD from available administrative data sources. The SPD is called the *Person Activity Register (PAR)*. Ireland is fortunate in that there is considerable usage of the Personal Public Service Number (PPSN) – which serves as the Person Identification Number in Ireland – across all public administration systems, along with the existence of a master register to validate basic information such as name, date of birth, gender and nationality. The master list of all PPSNs ever issued is maintained by the Social Welfare authorities in Ireland.

The availability of a high quality PPPSN on administrative data sources enables deterministic matching with a high degree of confidence. In practice, encrypted PPSNs are used for linking as a privacy protection measure. The master file of all PPSNs ever issued also provides a single source of truth for the key attributes, date of birth, gender and nationality, and, as such, eliminates any errors that may arise through domain misclassification between lists when linking or modelling data on these attributes.

The PAR has taken a different approach to building SPDs than that in many other countries. While the primary purpose of the PAR is to enumerate the population, the philosophy behind the PAR is one which seeks to minimise the number of different types of error to be addressed in compiling population estimates. For this reason, the PAR imposes strict criteria on which records to include. The PAR takes a *Signs-of-Life* approach and only uses the registration information from the master list of all PPSNs to provide consistent attribute information such as date of birth, nationality and gender. This Signs-of-Life approach can be summarised as only including persons

Table 2. Availability of Data Sources by year. Sources contributing to PAR, or List A, and sources proposed to be used as a second list, or list B, in a DSE approach to population estimation.

	Year					
	2011	2012	2013	2014	2015	2016
PAR - List A data sources						
Child Benefit	Y	Y	Y	Y	Y	Y
Early Childhood Care	N	Y	N	N	N	N
Post Primary Pupils	Y	Y	Y	Y	N	N
Higher Education Enrolments	Y	Y	Y	Y	Y	N
Further Education Awards	Y	Y	Y	Y	Y	Y
Employer Employee Tax Returns	Y	Y	Y	Y	Y	Y
Income Tax Returns (self-employed)	Y	Y	Y	Y	Y	N
Social Welfare	Y	Y	Y	Y	Y	Y
Public Health Benefits	N	N	Y	Y	Y	Y
State pension	Y	Y	Y	Y	Y	Y
List B data sources						
Driver Licence Dataset (DLD)	Y	Y	Y	Y	Y	Y
Quarterly National Household Survey	Y	Y	Y	Y	Y	Y

where there is evidence that they have engaged with the state and live in the State for a given reference year – this typically involves a financial transaction. The motivation for this approach is to eliminate or at least greatly reduce the need to deal with problems associated with *overcoverage*, *domain misclassification* (age, gender, nationality) and *linkage error*. Overcoverage is dealt with through choosing a suitable population concept and ensuring adherence to strict rules about whether to include a record or not, in practice, if in doubt chuck it out. Errors or inconsistencies with respect to domain misclassification are eliminated by using the original master list of all PPSNs as the truth for age, gender and nationality attributes. Unfortunately this master list of PPSNs does not contain high quality address information and as such there is no single source of truth to where a person resides leading to challenges when considering geographic distribution of the population. Linkage error is eliminated through the use of PPSN, and again if there is doubt over the quality of PPSNs in a group of records or data source it is not used. The only major remaining problem of any consequence to be dealt with is one of *undercoverage*.

The administrative data sources used to build the PAR involve sources across the full lifecycle, from the cradle to the grave, and include universal child benefit payments, health care payments, tax, education enrolments and awards (primary, secondary and tertiary) as well as social welfare payments such as State pension and unemployment benefit payments. Table 2 provides a summary of the availability of the different data sources by calendar year to be used in a DSE approach later described in section 3.

Not every data source is available each year and this should be noted. In practice new data sources will become available and some existing data sources may disappear or simply no longer be made available in a suitable form for the compilation of population estimates. Therefore, for the system to be effective over time it needs to be capable of

incorporating new data sources when they become available while at the same time be able to cope with the disappearance of existing data sources. We demonstrate a strategy to evaluate individual data sources in more detail later in Section 4.1.

In using these criteria, the PAR is considered to include persons that have been resident in the State at some time in the calendar year and have engaged with at least one Public Service. The population concept that underpins the PAR is the population of persons resident in the State that are entitled to engage with Public Services in the referenced calendar year. This allows other administrative data sources to be included at a later date if and when they become available. The criteria also only include persons considered to be resident in the State. However, there are slight differences when comparing this population concept to the usually resident population as enumerated in the traditional Census. The usual resident population concept typically refers to a point in time and has a requirement that persons should be resident or are intending to be resident for a period of at least 12 months.

In summary, the data sources underpinning the PAR provide broad coverage of the different stages of a persons life from the cradle to the grave. The PAR, taking a Signs-of-Life approach, contains records for only those people where there is strong evidence that the person was resident in the State for a given year. In particular, a relevant activity is admitted as evidence from the corresponding source only if the PPSN can be identified and verified. Direct counts from the PAR will therefore need to be adjusted for undercoverage errors if they are to be used as population estimates. Reducing the number of error types from four to one (undercoverage), as is proposed, creates a more favourable setting for compiling population estimates from administrative data. Capture-recapture methods can be used to estimate undercoverage.

2.2. Adjust SPD counts for Undercoverage to Obtain Population Estimates

In adjusting the PAR for undercoverage we reconsider the traditional approach for adjusting for Census undercount, which involves undertaking an undercoverage survey to generate recapture data and applying capture-recapture methods. These methods are introduced and covered in a number of texts (Bishop et al., 1975; Lohr, 2010; Rao, 2005).

Chao (2015) traces the use of capture-recapture ideas back to a 1786 paper by Pierre Simon LaPlace where it was used to estimate the population of France in 1802. An older example is identified where John Graunt used the idea to estimate the effect of plague on the population size of England around 1600.

Capture-recapture methodologies are often referred to as DSE in Official Statistics. The underlying assumptions, as described by Wolter (1986), are the traditional starting point when considering DSE to adjust for undercoverage. In section 3, we present a DSE model with more relaxed assumptions.

One administrative data source purposely not included in the PAR is the Irish Driver Licence database. A significant proportion of the adult population in Ireland hold a driving licence and are typically required to renew their licence every 10 years‡. However, renewal of licences can happen more often than every 10 years. For example, licence

‡More information on the rules with respect to the Irish Driver Licence System is available from http://www.ndls.ie , last accessed on 4th June, 2020.

categories such as learner driver licences and bus or truck licences will need to be renewed every 3 and 5 years respectively. Health grounds may also dictate a much shorter licence duration. Drivers are also allowed to apply for a replacement licence if they change address or if their licence is lost or stolen.

In addition to the PAR, the list of those persons that renewed their driving licences or applied for a new one in the relevant calendar year is proposed as a candidate for the other enumeration list, which will be denoted as the driving licence dataset (DLD). This means that the DLD in a given year only contains a subset of the persons that hold a drivers licence, those that renew or apply for a driver licence in that year. Historically, while the PPSN was requested it was not mandatory to provide it to obatin a new or renewed licence. However, since 2013 the provision of a verified PPSN has become mandatory. Again, a person is included in the DLD provided only the PPSN is identified and verified.

Any person normally resident in the State is allowed to apply for, or renew, an Irish Driver licence subject to the usual age restrictions. A person is considered normally resident if, because of personal or occupational ties, they live in the State for more than 185 days in a given year. Since these persons are entitled to Public Services in the same year, the DLD and the PAR are considered as subsets of the same underlying target population. The PAR will contain persons that have never held a driving licence while the DLD will contain persons that renew or apply for a driver licence that may not engage with other public services. For example, a healthy person who is dependent on another person or who as unofficial sources of income may not have a requirement to engage with any of the public services underpinning the PAR yet may renew or apply for a drivers licence. In summary, we consider PAR and DLD as list inputs to DSE methods in compiling population estimates. DSE methods include an estimate of persons in the population excluded by both lists.

Blocking both the PAR and the DLD by single year of age, nationality grouping and gender, a DSE will be calculated within each block. Blocking in this manner facilitates a disaggregation of population estimates which is comparable to the demographic component method. Moreover, this allows for likely different propensities to hold a driving licence across the blocks, with respect to the *homogeneous capture* assumption that will be discussed in Section 3 below.

3. Methodology - the PECADO toolkit

3.1. Overview

Let the PAR be list A, and let the DLD be list B. The traditional assumptions of DSE used are typically those presented by Wolter (1986), which include 8 assumptions. In the formulation of this paper, there are only three assumptions required to compile the DSE and an additional *fourth* assumption for variance estimation (Zhang and Dunne, 2018; Zhang, 2019) The assumptions are

- *Matching:* There is no error when matching records between list A and list B.
- *No erroneous records:* There are no records from outside the population, there are no duplicate records or incorrectly identified records.

- *Homogeneous capture (List B):* Every unit in the population has an equal chance of being included in list B.
- *Independent capture (List B):* Whether one population unit is included in list B is independent of whether any other population unit is included in list B.

This relaxation of assumptions allows for a far more flexible application of DSE. Zhang (2019) discusses in detail the relationship between these 4 assumptions and Wolter's eight assumptions. In particular, as explained below, Wolter's assumption of (causal) independence and closed population can be removed in our DSE setup and the assumption of equal catchability only needs to hold for one of the two lists. The 4 required assumptions are explained in Section 3.2. Exploring the behaviour of this DSE formulation in the presence of erroneous records provides us with an extension of the DSE method that allows users to deal with the presence of erroneous records. This extension is presented in Section 3.3. Finally, a consideration of the violation of the *homogeneous capture* assumption in Section 3.4 gives an understanding that even if this assumption is violated the DSE estimate may still be valid and presents a test that evaluates whether a potential violation of the assumption across groups is significant.

3.2. DSE

Let N be the unknown size of the target population, denoted by U. Let A be the first list of size x. Suppose list A is subject to undercoverage so that $x < N$ and $A \subset U$. Let B be the second list of size n and also subject to undercoverage so that $n < N$ and $B \subset U$. This requires the assumption of *no erroneous record* in $A \cup B$.

Suppose the records in list A and list B can be linked in an error free manner and doing so will provide the matched list AB with m records common to both list A and list B. This requires the *matching* assumption.

Let $\delta_{iB} = 1$ if $i \in B$, noting $B \subset U$, and 0 otherwise. We assume that the probability $P(\delta_{iB} = 1) = \pi$ is a constant across $i \in U$, which is the *homogeneous capture* assumption of list B. It is the starting point of the development of the estimator. Heterogeneous capture can be accommodated through post-stratification which requires that the homogeneous capture assumption holds within each post-stratum.

Given the assumption of homogeneous capture, we have

$$E[n] = N\pi$$

Moreover, let $\delta_{iA} = 1$ if $i \in A$, noting $A \subset U$, and 0 otherwise. For any $i \in U$, we have

$$P(\delta_{iB} = 1) = P(\delta_{iB} = 1|\delta_{iA} = 1) = P(\delta_{iB} = 1|\delta_{iA} = 0) = \pi$$

Notice that here we consider $\boldsymbol{\delta}_A = (\delta_{1A}, ..., \delta_{NA})$ as fixed constants, where $\sum_{i \in U} \delta_{iA} = x$. The above equalities are therefore merely consequences of the assumption of homogeneous capture, and do *not* formally amount to an assumption of independence between δ_{iA} and δ_{iB}.

Provided the assumptions of homogeneous capture and matching hold, we have:

$$E[m|\boldsymbol{\delta}_A] = x\pi$$

which is the expectation of the number of records in list AB ($A \cap B$) on applying the constant capture probability π to the x records in list A with $\delta_{iA} = 1$. Replacing $E[n]$ by n and $E[m|\boldsymbol{\delta}_A]$ by m, we obtain a method-of-moment estimator, given by

$$\hat{N} = nx/m \tag{1}$$

where x is a fixed constant and only (n, m) vary due to the randomness of list B.

The variance of $\hat{N}$ is obtained by means of the additional assumption of *independent capture* (of List B), such that $V[n] = N\pi(1-\pi)$ and $V[m] = x\pi(1-\pi)$. Writing $n = m + n_{A^c}$ where n_{A^c} is the number of population units that are not in list A but are enumerated in list B, we have $Cov[n, m] = Cov[m + n_{A^c}, m] = V[m]$. Thus, by the linearisation technique, we obtain

$$V\left[\hat{N}\right] \approx \frac{x^2}{E[m]^2}\left(V[n] - \frac{2E[n]}{E[m]}Cov[n,m] + \frac{E[n]^2}{E[m]^2}V[m]\right)$$

$$= N(\frac{1}{\pi} - 1)\left(\frac{N}{x} - 1\right)$$

Replacing N by xn/m and π by m/x, we obtain a variance estimator

$$\hat{v} = \widehat{V}\left[\hat{N}\right] = \frac{n(n-m)x(x-m)}{m^3} \tag{2}$$

Notice that this is the same variance estimate as that of the standard DSE described in the text book of Bishop et al. (1975), where both lists are treated as random and independent (i.e., the probability of being in both list A and list B equals the probability of being in list A multiplied by the probability of being in list B).

The relaxing of the assumptions in this derivation is important. It means that the DSE can now be applied in many more scenarios and in particular to a scenario where list A is derived from administrative data sources and the argument or assertion that all the assumptions described by Wolter (1986) need to apply is weak.

Chao et al. (2008) also explores this concept of independence between the two lists and shows that *equal-catchability* or *homogeneous capture* for the second sample will suffice. They note that some may state this assumption as one sample being a representative sample or simple random sample. They also discuss the importance of this finding in the context of the Census undercount application. For instance, if all individuals in the population have equal or similar probability of being counted in the Census then the Census can be considered a simple random sample and as such a coverage survey can have heterogeneity in the capture rates.

3.3. Trimmed DSE - dealing with erroneous records in list A

3.3.1. Ideal DSE, given erroneous enumeration

Suppose we relax the assumption that there are no erroneous records and develop the estimator under the same assumptions otherwise.

Let N again be the unknown size of the target population, denoted by U. Let list A be of size x. Suppose list A contains r erroneous records, i.e. the size of $\{i; i \in A \text{ and } i \notin U\}$.

Suppose list A is subject to under-counting as well, so that $x - r < N$. Let B be the second list that is of size n. Suppose list B is subject to *only* under-counting, so that $n < N$, but there are *no* erroneous records in B.

Again suppose the records in lists A and B can be linked to each other in an error-free manner which gives rise to the matched list AB with m records.

Again let $\delta_{iB} = 1$ if $i \in B$, noting $B \subset U$, and 0 otherwise. We assume that the probability $P(\delta_{iB} = 1) = \pi$ is a constant across $i \in U$.

Thus, allowing erroneous records in A but retaining the other assumptions, we obtain

$$E[n] = N\pi \qquad \text{and} \qquad E[m|\boldsymbol{\delta}_A] = (x - r)\pi$$

where the latter is the expectation of the number of records in list AB on applying the constant capture probability π to the $x - r$ in-scope records in list A. Replacing $E(n)$ by n and $E[m|\boldsymbol{\delta}_A]$ by m, we obtain an *ideal* method-of-moment estimator, insofar as r is unobserved, given by

$$\tilde{N} = \frac{n(x - r)}{m} \tag{3}$$

Meanwhile, let the naïve DSE, which ignores the erroneous records in list A altogether, be given by

$$\dot{N} = nx/m$$

It follows immediately that $\dot{N}$ can be expected to *over-estimate* N, since $n(x - r)/m < nx/m$ for any $r > 0$.

Notice, that provided suitable estimation methods such as the Trimmed DSE described below exist, we no longer need to assume that the target population is closed for both lists, as long as it is possible to correctly identify the population units in list B, and the matching between A and B is error-free. One only needs a particular version of $\boldsymbol{\delta}_A$ that is matched to list B, even if $\boldsymbol{\delta}_A$ itself can change due to the updating of list A over time. The units with $\delta_{iA} = 1$ are simply the 'marks' that allow the estimation of the capture probability π of list B.

3.3.2. Trimmed DSE

The estimator (3) is hypothetical because r, the number of erroneous records in list A, is unknown. But one *can* (a) trim some records from list A which are suspected of being erroneous, (b) match the trimmed list A to list B and, then, (c) calculate the new DSE with this new trimmed list A and the match it generates (Zhang and Dunne, 2018).

This yields what we call the *trimmed DSE*, given by

$$\hat{N}_k = n\frac{x - k}{m - k_1} \tag{4}$$

where k is the number of trimmed records in list A, and k_1 is the number of records among them that can be matched to list B. Notice that, provided list B has only under-count, the k_1 records are indeed not erroneous, whereas the remaining $k - k_1$ records may or may not be erroneous.

The trimmed DSE can be compiled under the *same* assumptions as those for the ideal DSE, as per equation (3), *regardless* of how systematic the trimming is in removing

records from list A. Potential systematic undercoverage of list A does not matter to start with. For instance, had one trimmed all the people between 20 and 25 years old in list A, the trimmed DSE, $\hat{N}_k$, would have remained a valid estimate provided all the erroneous records had been removed in this way. Zwane et al. (2004), in their investigations with the Multiple Systems Estimator, also conclude that the Peterson-Lincoln estimator is still valid if one group is missing from a list provided the second list is a sample across all groups with homogeneous capture probabilities.

As shown above, the naïve DSE, which can now be written as $\hat{N}_0$ with $k = 0$, is expected to over-estimate N. The following results from Zhang and Dunne (2018); Dunne (2020) are useful in applying TDSE methods to evaluate and trim parts of list A where erroneous records may exist.

i) If $k_1/m < k/x$, then $\hat{N}_k < \hat{N}_0$. There is evidence of erroneous records in the trimmed element of list A.

ii) If $k_1/m = k/x$, then $\hat{N}_k = \hat{N}_0$. There is no evidence of erroneous records in the trimmed element of list A.

iii) If $k_1/m > k/x$, then $\hat{N}_k > \hat{N}_0$. There is evidence of erroneous records remaining in the untrimmed element of list A.

iv) If $k < r$, then $\tilde{N} < \hat{N}_k$. The trimmed estimate cannot remove all bias due to erroneous records when $k < r$.

v) If all the r erroneous records are among the k trimmed ones, then $\widehat{E}[\hat{N}_k] = \tilde{N}$.

To summarise, as long as one is able to trim the erroneous records in list A more effectively than when randomly trimming records, the TDSE (4) can be expected to reduce the bias of the naïve DSE and move it closer to the ideal DSE (3). If the trimming succeeds in removing all erroneous records, the expectation of the TDSE will become approximately the same as the ideal DSE.

When it comes to variance estimation, consider first the ideal estimator $\tilde{N}_k = \tilde{x}n/m$ where $\tilde{x} = x - r$. The variance of the ideal TDSE $\tilde{N}$ and its estimator are given by those of $\hat{N}$ at (1), on replacing x by $\tilde{x}$. Rewrite the TDSE as $\hat{N}_k = x_k n/m_k$, where $x_k = x - k$ and $m_k = m - k_1$. For variance estimation, one needs the number of remaining erroneous records among the trimmed list A with x_k records, which is not known. As a practical remedy, we assume the adopted TDSE is successful to the extent that $E(\hat{N}_k) \approx N$, i.e. all the x_k records belong to the population, so that a variance estimator of $\hat{N}_k$ can be given by

$$v_k = \widehat{V}\left[\hat{N}_k\right] = \frac{n(n - m_k)x_k(x_k - m_k)}{m_k^3} \tag{5}$$

If we consider an ineffective trimming strategy where k_1/k is large and relatively many records need to be trimmed before all erroneous records are removed from list A, equation (5) shows that there is a danger that the trimmed estimate will become unstable due to high variance (m_k reduces much faster than x_k).

3.4. Testing the homogeneous capture assumption

3.4.1. Exploring impact on estimates due to heterogeneity in capture rates

There are a number of papers in the literature, which deal with heterogeneity of capture rates over subgroups. The matter can be approached as selecting a log-linear model for a contingency table, where two of the dimensions correspond to the two capture-recapture lists and the others are introduced for the available covariates (or subgroup indicators). It has also been considered to use an unobservable (latent) categorical variable to identify groups with heterogeneous capture probabilities (Stanghellini and van der Heijden, 2004), in situations involving three or more enumeration lists. Here, we consider the impact of a violation of the homogeneous capture assumption with respect to post-stratification that is commonly applied for DSE (Section 3.2).

To consider the impact of a violation of the *homogeneous capture assumption*, or heterogeneity in the capture rates over the subgroups, we take as our starting point the setup in Section 3.2.

We start with a population U of size N. We have a list A of size x where we know $A \subset U$ and we have a list B of size n where we make the assumption that each person in U has an equal chance $E(n)/N$ of being included in list B. The DSE is $\hat{N} = nx/m$ where m is the size of list AB, the match between lists A and B.

We now consider a partition of our population U based on covariate information, common to both lists A and B, into two subgroups U_1 and U_2 where $U_1 \cap U_2 = \emptyset$ and $U_1 \cup U_2 = U$. We also let N_1 and N_2 denote the sizes of the two partitions U_1 and U_2 respectively, noting $N_1 + N_2 = N$. Using the covariate information, we can now partition list A into lists A_1 and A_2 where $A_1 \subset U_1$ and $A_2 \subset U_2$. Similarly, we partition list B into lists B_1 and B_2 where $B_1 \subset U_1$ and $B_2 \subset U_2$. We let x_1, n_1, m_1, x_2, n_2 and m_2 denote the list sizes for lists A_1, B_1, $A_1 \cap B_1$, A_2, B_2 and $A_2 \cap B_2$ respectively. If there is a suspicion of heterogeneity in capture rates between U_1 and U_2, that is $E(n_1)/N_1 \neq E(n_2)/N_2$ then a two-part (stratified) DSE estimator is more sensible, which is given by

$$\hat{N}' = \hat{N}_1 + \hat{N}_2 = \frac{n_1 x_1}{m_1} + \frac{n_2 x_2}{m_2} \tag{6}$$

Since $n = n_1 + n_2$, $x = x_1 + x_2$ and $m = m_1 + m_2$, we can rewrite $\hat{N}$ above as

$$\hat{N} = \frac{nx}{m} = \frac{(n_1 + n_2)(x_1 + x_2)}{(m_1 + m_2)} \tag{7}$$

The estimator in equation (6) allows for differences in the capture rate for list B between the two parts while the estimator in equation (7) requires the capture rate to be the same between the two parts. Now, we can explore the difference between the two estimators to consider the impact of any violation of the *homogeneous capture assumption* for the population U in list B.

Manipulating equations (6) and (7) we can express their difference as

$$D = \hat{N} - \hat{N}' = \left(\frac{n_1}{\hat{N}_1} - \frac{n_2}{\hat{N}_2}\right)\left(\frac{x_2}{\hat{N}_2} - \frac{x_1}{\hat{N}_1}\right)\frac{\hat{N}_1 \hat{N}_2}{m} \tag{8}$$

Details of this manipulation are available in Dunne (2020). It is clear form (8) that $D = 0$ as long as $x_2/\hat{N}_2 = x_1/\hat{N}_1$ even when the list-B capture probability varies across

the two parts. In other words, heterogeneous capture of list B on its own does not necessarily cause a large bias of $\hat{N}$.

3.4.2. Is the difference due to heterogeneity in capture rates significant?
Consider the general case of heterogeneity in capture probabilities across subgroups indexed by h. We propose to test the null hypothesis $H_0 : E(D) = 0$, where

$$D \coloneqq \hat{N} - \hat{N}' = \frac{\sum_h n_h \sum_h x_h}{\sum_h m_h} - \sum_h \left(\frac{n_h x_h}{m_h} \right)$$

with variance

$$V(D) = V(\hat{N}) + V(\hat{N}') - 2Cov(\hat{N}, \hat{N}')$$

Under the null hypothesis, we have

$$\hat{V}(\hat{N}) = \frac{n(n-m)x(x-m)}{m^3}$$

for $n = \sum_h n_h$, $m = \sum_h m_h$ and $x = \sum_h x_h$, and

$$V(\hat{N}') = \sum_h \left(\frac{n_h(n_h - m_h)}{x_h} \frac{(1-\hat{\pi})}{\hat{\pi}^3} \right)$$

where m_h/x_h in each subgroup can be replaced by $x_h\hat{\pi}$. A detailed derivation concerning $Cov(\hat{N}, \tilde{N})$ is given in Dunne (2020), which yields an estimate

$$\widehat{Cov}(\hat{N}, \hat{N}') = \frac{(1-\hat{\pi})}{\hat{\pi}^2} \left(\frac{n^2}{m} - n \right)$$

We obtain thus a test statistic $Z = D/\sqrt{\hat{V}(D)} \sim N(0,1)$ for H_0.

For an illustration, suppose list sizes $(x_h, n_h, m_h) = (50, 40, 20), (100, 90, 30)$ and $(150, 120, 90)$ for three subgroups, such that $\hat{\pi}_h = (0.4, 0.3, 0.6)$. We obtain the population size estimates $\hat{N} = 536$ and $\hat{N}' = 600$, as well as $D = -64$, $\hat{V}(D) = 12$ and the test statistic $Z = -18.48$, which indicates highly significant bias due to heterogeneous capture across the subgroups if the heterogeneity is ignored in estimation.

4. PECADO - a robust system of population estimates

We implement and evaluate the system of population estimates proposed in Section 2 using the toolkit presented in Section 3.

To recap, the system of population estimates can be summarised as follows. We build the PAR using a Signs-of-Life approach to act as our list A, which aims to remove the erroneous records in the underlying sources. We then use the DLD as list B in a capture-recapture system to adjust for undercount of the PAR. The linkage between A and B is based on the PPSN available across the administrative sources, which satisfies the matching assumption. DSE is applied together with blocking by gender, year of age and

nationality grouping, where the attributes for the blocking variables are all taken from the same master register to ensure no domain incoherence between lists. Blocking both reduces the impact of heterogeneous capture probabilities of the DLD and facilitates disaggregation comparable to the demographic component method.

Notwithstanding the theoretical reasoning, there may always be weaknesses in the underlying assumptions in practice. Below we describe first our strategy to the evaluation of the relevant data sources (Table 2) and the implementation of the PECADO system. Next, we present the obtained population estimates for 2011 - 2016 and discuss the differences to the Census counts 2011 and 2016. Finally, we apply the PECADO toolkit to investigate the potential undercoverage errors of Census 2016, demonstrating how this can be accomplished using the methods described in Section 3 in the absence of an undercoverage survey, as it is the case in Ireland.

4.1. Evaluation and implementation strategy

The methods developed in Section 3 should be considered in conjunction with the underlying data sources following an overall evaluation and implementation strategy, which involves a number of iterative steps. In our case, where linkage error is negligible due to the PPSN, the strategy is focused on the potential erroneous records in the lists and the choice of blocking for the DSE, as reflected in the high-level process representation (Figure 1).

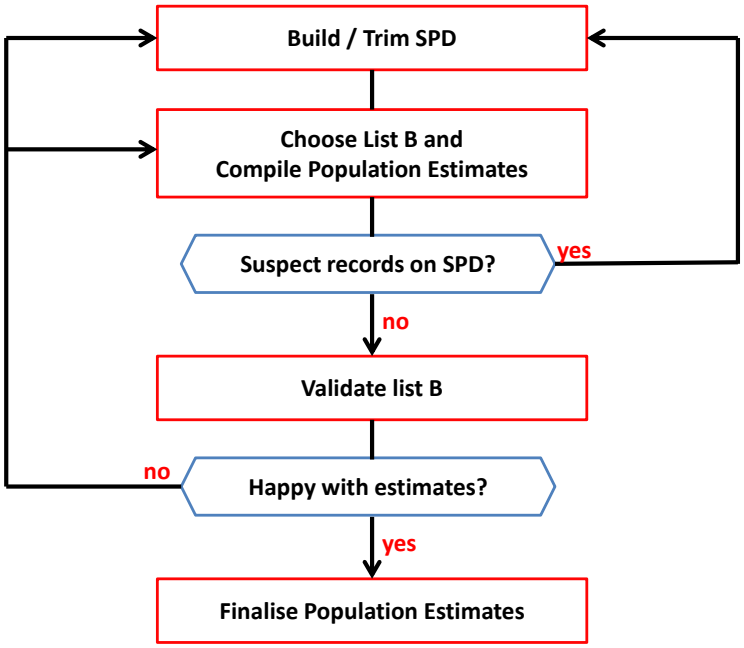

Fig. 1. High level process map of evaluation and implementation strategy.

To begin with, we evaluate each data source (Table 2) using the TDSE methodology. The trimming steps involve dropping each underlying data source in turn before rebuilding the list A and compiling the population estimates. Population pyramids of estimates and confidence intervals (e.g. Figure 2) are used to evidence whether a trimmed data source suffers from erroneous records or not. If a data source is deemed to be unsuitable due to erroneous records it is then dropped from the list A. The population pyramids can also provide an insight into the importance of each data source to the precision of the estimates through a comparison of confidence intervals. A trimming step could also include dropping two underlying data sources in one step if there is a suspicion that the two data sources are similar with respect to erroneous records contained within.

Next, it can be seen in Figure 1 that list B (hence, the DSE) is always revalidated after trimming of the list A and compilation of population estimates. The reason for this is that, if the list A contains erroneous records, this in itself may have an impact on validating list B. In particular, we can evaluate the chosen dataset (DLD in this instance) as a suitable candidate for list B with the following approach:

- identify an alternative data source that can be used as list B, such as a large household survey
- use this alternative data source to create a new set of population estimates
- compare this alternative set of population estimates with the original set to see if they are consistent.

If the two sets of population estimates are consistent with each other, then both data sources used as list B can be considered. If not, this leads to the conclusion that one or both of the two data sources violate the homogeneous capture assumption. The compatibility of the two sets of estimates again can be evaluated by plotting the estimates along with confidence intervals on a population pyramid type plot.

As an alternative candidate for list B we use the Quarterly National Household Survey, which employs an equal-probability sampling design. The survey returns are linked to administrative data sources using Personal Identification Information such as name and date of birth to obtain the relevant PPSN where possible. Additional assumptions are required to use the Quarterly National Household Survey as list B – both the non-response and unlinked units are missing-at-random. The list size is also much smaller than the DLD and the variance of the corresponding DSE is inflated due to the two-stage design, which induces a correlation among the capture of the same household members. Despite these complicating factors, alternative estimates using the Quarterly National Household Survey did help to inform the final decision that the DLD can be accepted as list B. The two sets of estimates are in broad agreement with each other, and the survey does not make a better list B not least because of the unavoidable nonresponse errors; more details are given in the online reference material.

As we have iterated through the steps in the evaluation and implementation strategy and validated the DLD as list B, it is worth noting the following decisions regarding the preparation of both the lists.

- One data source that contained a proxy for state pension recipients was dropped as there was evidence of erroneous records that were biasing the population estimates

Table 3. Population estimates compiled from administrative data sources, 2011 to 2016.

Year	2011	2012	2013	2014	2015	2016
List A	4,397,770	4,424,370	4,533,430	4,541,630	4,611,800	4,473,900
Coverage (%)	91	92	93	92	92	89
List B	422,680	507,030	468,870	378,100	466,610	539,200
Match AB	376,950	452,730	425,130	341,230	422,070	462,330
Estimate	4,811,020	4,828,990	4,896,230	4,925,380	4,992,260	5,038,640
CV (%)	0.05	0.05	0.04	0.05	0.04	0.05

in the upper age categories.

- The Public Health Benefits data source for 2013 were aged appropriately and included as a data source in the PAR for 2012 and 2011 to counter the poor coverage in these years for the older age categories (as a result of dropping pension proxy). This is justified on the basis that the older age categories are not considered to be affected by migration in any significant way.

- Only the DLD was used as list B. This is justified on the basis that we assume no significant undercoverage in the under 18 age category and that this part of the population does not need to be adjusted (An examination of Figure 2 and Table 5 would supports this assumption). There is sufficient coverage of the retirement age category by the PAR, i.e. $x/\hat{N} \doteq 1$ that any violation of the *homogeneous capture assumption* with respect to DLD for these categories will only have a minor impact on the estimate in terms of bias.

- Workers with less than 20 weeks employment recorded are removed from Employer Employee Tax Returns before it is included in the PAR. This has the effect of tuning the estimate, such that the underlying population concept equates to that of an *annual resident population* (Lanzieri, 2013) and is better aligned to the commonly used concept of usual residence (12 months, intended or actual), by excluding temporary or migrant workers who may come and work for a period of 20 weeks or less.

4.2. PECADO population estimates

The population estimates, along with precision estimates, for years 2011 to 2016 are presented in Table 3. Coverage of the SPD for 2016 drops slightly as the income tax returns for the self employed and the Higher Education Enrolment data were unavailable for 2016 at time of compilation (see Table 2).

A comparison of the population estimates and Census usual resident counts by gender for 2011 and 2016 are provided in Table 4 and this comparison, broken down by age is presented using population pyramids in Figure 2 for 2016. The gap between the new population estimate and the Census usual resident count widens from 5.2% to 6.3% between 2011 and 2016. The gap is wider for males than for females. When differences are explored using the population pyramids we see that the biggest differences between the population estimates and Census usually resident count occurs for young adult males between the ages of 20 and 40 years old.

Table 4. Comparison of PECADO Population Estimates with Census usual resident counts by gender, 2011 and 2016.

	Population Estimate	Census	Difference Total	(%)
2011				
Both Sexes	4,811,020	4,574,890	236,130	5.2
Male	2,421,310	2,270,510	150,800	6.6
Female	2,389,710	2,304,390	85,320	3.7
2016				
Both Sexes	5,038,640	4,739,600	299,040	6.3
Male	2,539,120	2,346,550	192,570	8.2
Female	2,499,520	2,393,050	106,470	4.4

There are mainly 4 possible explanations for the difference between the population estimates and the Census (usual resident) count.

The first explanation relates to the underlying population concept and dynamics. The Census counts those usually resident in the state on Census night. A person is considered usually resident if they have been living in the state for 12 months or more or are currently resident with the intention of being resident for 12 months or more. The population estimates are based on those resident in the State for a significant period at any given point in the calendar year. The signs of life for inclusion on the PAR have been tuned to only include those where the sign of life is indicative that the person is or will be resident for a significant period. For this reason, signs of life related to short periods of work ($<$ 20 weeks) have been removed. The primary difference between the two is that the Census count represents a snapshot on a specific night while the population estimate can relate to any night in the calendar year. This would imply that to equate the Census count to the population estimate, emigration plus deaths prior to Census night in the calendar year and immigration plus births subsequent to Census night in the calendar year must be added for usual residents. Table 1 gives an estimate of 230,000 (95,000 inflows and 145,000 outflows) for gross population flows in a 12 month period, or approximately 4.8% of the Census 2016 population count. One can conjecture a difference of approximately $130{,}000 \approx (1/3)95,000 + (2/3)145,000$ or 2.8% if we assume population flows are evenly distributed throughout the calendar year.

The second explanation is the existence of yet to be identified erroneous records on either the PAR or DLD. While considerable scrutiny has already been given to the underlying data sources contributing to the PAR, and problematic data sources removed, we have to acknowledge that it is not impossible that there may still be erroneous records on the PAR. Since 2013, the rules governing renewing a driver licence have become far stricter in terms of identification and therefore it is reasonable to assume that there are only a negligible amount of erroneous records in the DLD. The DLD has also been validated (Section 4.1) as a list B data source. While this validation focussed on the homogeneous capture assumption, the validation would not have been successful had DLD contained a significant quantity of erroneous records.

A third explanation is that a violation of the homogeneous capture assumption would lead to bias in the population estimates. While earlier analysis (see Section 4.1) gener-

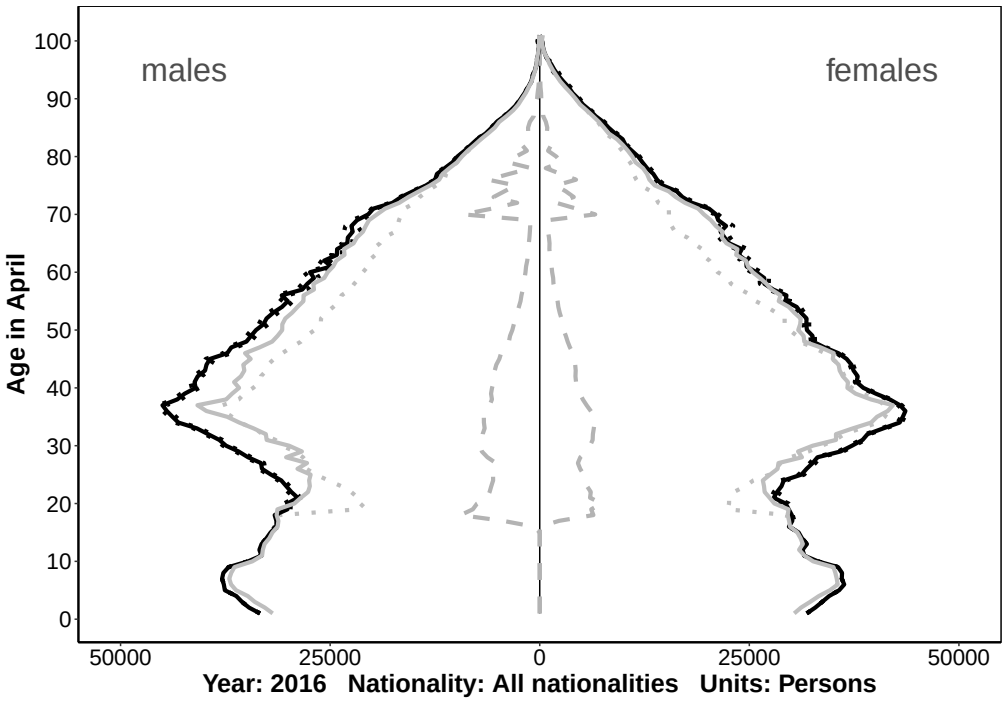

Fig. 2. Comparison of population estimates and Census counts by age and sex, 2016. Black continuous line with dots - $\hat{N}$ with 95% confidence intervals, grey continuous line - Census usually resident count , grey dotted line - list A counts, grey dashed line - list B counts

ally validated the choice of blocking for the DSE, there were some indications of small violation of the homogeneous capture assumption for females in the older age category. Specifically, first, when the data source containing state pension records is removed, the population estimate drops below that of the Census; then, when DLD is replaced with QNHS, the population estimate moves back above the Census, albeit with a larger confidence interval.

The fourth explanation may be that the Census has an undercount with respect to the usual resident population. As mentioned earlier, Ireland does not conduct a coverage survey and assumes the Census operations are sufficiently robust to ensure any coverage issues can be ignored. Statistics New Zealand reported the net Census undercount as 2.4% in the 2013 Census (Statistics New Zealand, 2014). For the 2018 Census, the Census undercount has been estimated at just over 10% before administrative records are used to assist in the completion of the final dataset (Bycroft et al., 2018). Northern Ireland recorded an undercount of 4.7% in the 2001 Census (Abbott, 2009) and a similar undercount in 2011 after additional active persons on administrative data sources were included (NISRA, 2015). The undercount was significantly greater for young adults. ONS estimated the Census 2011 undercount to be about 6% overall, 7% for males and 5% for females. Given these findings in countries broadly similar to Ireland, it is reasonable to consider the possibility of undercount in the Irish Census. In Section 4.3 next, we

show how this can be investigated using the PECADO toolkit, despite the lack of a traditional undercoverage survey.

4.3. Exploring potential Census undercount

We combine Social Welfare Signs of Life data and the Census in a DSE set-up. In terms of the assumptions explained in Section 3.2, we consider the Census as list B, which is assumed to have homogeneous capture (subject to blocking). Using Social Welfare as list A saves the cost of conducting a separate undercoverage survey in a traditional manner. The resulting population estimate implies then an estimate of Census undercount.

In April 2016, CSO enumerated 4.74 million persons. Subsequent to the Census, CSO undertook a matching exercise to identify a link between persons enumerated in the Census and administrative records associated with that person. The purpose of creating the link was to enable enhanced statistical products (e.g. Household Income) through linking the Census data with administrative data.

Relevant privacy safeguards were deployed in creating the linkage framework between the Census dataset and administrative data sources. Subsequent quality interrogations and validation exercises show the linking to be of high quality. There was a number of records on the Census dataset for which it was not possible to identify, with sufficient confidence, an associated link with administrative data. It is assumed that the reason for being unable to match this subset is that they contained poor quality name information and/or incorrect date of birth information.

Limiting the Census dataset to only those records that could be linked to administrative records (or an official PPSN), resulted in a Census dataset of 4.27 million records. We will refer to this version of the Census dataset as the *trimmed* Census dataset.

List A contains the records for each person receiving a Social Welfare payment in April 2016, which is closer to the Census concept (point in time) than the the PAR (any part of the reference year). This list A also contains the same encrypted identifier key as on the *trimmed* Census dataset. Given that the Government Department responsible for Social Welfare payments invests significant resources in ensuring customers are entitled to a payment - in other words, appropriate procedures are in place to authenticate these persons as living in the State and having an entitlement to a payment from the respective scheme, it is reasonable to accept the persons on list A as usually resident around the time of Census night.

The estimates are presented in Table 5. Due to the low coverage of school age by the Social Welfare data source, the DSE is applied to the age group 20 and over, giving the so-called direct population estimate in the table. The final population estimate of 5,001,700 persons is obtained from adding to this the PECADO estimate for the under 20 year age group. Comparing this figure with the Census count of 4,739,600 implies an estimate of Census undercount of approximately 262,100 or 5.2%. Moreover, a closer inspection of the population pyramids shows that the undercoverage is primarily in the age categories 20 to 40, which is the part of the population where migrating is highest.

For the Northern Ireland 2011 Census (NISRA, 2015), just under 92% of usual residents were included in an adequately completed questionnaire, a further 4% were captured through using administrative data with the remaining about 5% being derived through a coverage assessment and adjustment process. In this context, an estimate of

Table 5. Comparison of Census Usual Resident counts, Preliminary Census Usual Resident estimates adjusted for undercoverage and PECADO population estimates by age group, 2016.

	Age group		
	Under 20	20 and over	All ages
Census	1,306,700	3,432,900	4,739,600
Trimmed Census	1,193,000	3,074,900	4,267,900
Social Welfare April	15,400	1,397,300	1,412,700
Match	13,300	1,160,000	1,173,300
Population Estimate (Direct)		3,668,300	
PECADO Population Estimate	1,333,400	3,705,300	5,038,600
Population Estimate (Final)	1,333,400	3,668,300	5,001,700
Implied Census Under-coverage (%)	2.0	6.4	5.2

undercoverage at 5% is relatively good. However, it does suggest that CSO, Ireland can not rely on accepting at face value that the current implementation of the traditional Census model can eliminate undercoverage. Some form of assessment or adjustment is required and planned for as part of the next Census in 2022.

5. Final remarks

This paper presents a robust system of population estimates for Ireland, based solely on administrative data sources and in the absence of a Central Population Register. Special attention is given to three aspects of the development of this PECADO system: (I) a Signs-of-Life approach to the preparation of data sets as the basis of estimation, and the supporting master register and linkage key; (II) innovations of the capture-recapture methodology, which are both necessary and useful for adaptions from the traditional setting of Census and Census Coverage Survey; and (III) a viable strategy of implementation and validation of the PECADO system. Applying the same ideas and techniques to the Census 2016 enumeration allows us to explore the potential census undercoverage error without the need of a separate undercoverage survey.

The presence of an identification number is not a pre-requisite for linking administrative datasets to compile an SPD. Some countries are already linking data sources in the absence of official identification numbers, New Zealand (Statistics New Zealand, 2014) and United Kingdom (ONS UK, 2017) to name two. The linking strategy typically involves creating some composite key of some or all of name, date of birth, address, linked family members. The applications in this paper have not only focused on linking different administrative datasets to each other with an official identification number, but have also linked survey/census datasets to administrative datasets by first linking the survey/census datasets to the official list of identification numbers. The linking strategy adopted in linking survey/census data to administrative data was one of minimising or eliminating false positives/negatives by not assigning official identification numbers where there is any doubt. The records that weren't linked are removed from the lists in question and assumed missing at random. The PECADO toolkit has applications in situations where no identification numbers exist particularly when careful consideration is given to how to eliminate linkage error and or how to deal with linkage error in any final estimates. There may also be opportunities to further evaluate the PECADO toolkit for

estimating Census undercoverage in settings where a traditional Census undercoverage survey has already been undertaken.

The paper concludes that the PECADO system can provide robust annual population estimates by gender and single year of age for Ireland and is a solid foundation on which to build a modern system of annual census-like population estimates. The underlying toolkit may, in whole or part, also provide a valuable contribution for other countries when modernising or enhancing their systems of population estimates.

In considering further developments of the PECADO system, three directions are noted.

First, the PECADO system adopts an annual resident population concept which is more congenial to the available data but can then be easily tuned or adjusted to other concepts by adjusting for estimated migration numbers. As many countries are currently engaged in their respect census transformation programmes, the harmonisation of existing and emerging population concepts calls for greater effort and cooperation internationally, as e.g. discussed by Lanzieri (2013), in order to revise the relevant recommendations and guidelines that can facilitate comparisons over time and across countries. The PECADO system should take active part in this development.

Second, a key advantage the PECADO system holds over the demographic component method is that it does not depend on first estimating the migration components of change. This in turn implies the PECADO system is not dependent on a large Census operation every 5 or 10 years for recalibration, which also triggers a significant revision of the many other sectoral statistics that depend on population estimates as an input. In this regard, Dunne (2020) demonstrates how the PECADO system can be extended to provide estimates of gross population flows in a different approach, where one first estimates the Stayers (those persons that are members of the population for two consecutive years, year 1 and 2) before differencing this estimate of Stayers from the population estimate in year 1 (to obtain an estimate of outflows in year 1) and the population estimate in year 2 (to obtain an estimate of inflows in year 2) to obtain estimates of gross population flows. The approach has the added benefit of ensuring coherency between the estimates of stocks and flows.

Third, a key consideration for annual census-like population estimates is to be able to disaggregate population estimates to detailed geography. Differing location attributes between list A and list B presents a challenge in the PECADO system and DSE generally. To meet this challenge the authors are evaluating another extension of the estimation framework, Calibrated Dual System Estimation, where the inter-locality over- and undercounts caused by the problem of misplacement are adjusted simultaneously.

References

Abbott, O. (2009). 2011 UK Census coverage assessment and adjustment methodology. *Population trends*, 137(137):25–32.

Bishop, Y., Feinberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis*. Springer.

Bycroft, C., Connolly, K., and Quinn, A. (2018). Transcript : 2018 Census Technical Seminar.

Chao, A. (2015). Capture-Recapture for Human Populations. *Wiley StatsRef: Statistics Reference Online*, pages 1–16.

Chao, A., Pan, H. Y., and Chiang, S. C. (2008). The Petersen - Lincoln Estimator and its Extension to Estimate the Size of a Shared Population. *Biometrical Journal*, 50(6):957–970.

Dunne, J. (2020). *The Irish PECADO project: Population Estimates Compiled from Administrative Data Only*. PhD thesis, University of Southampton.

Dunne, J. and Graham, P. (2019). New Population Estimation Methods : New Zealand and Ireland. In *ISI World Statistics Congress 2019*, number August, Kuala Lumpur.

EUROSTAT (2003). *Demographic Statistics: Definitions and Methods of Collection in 31 European Countries*. European Communities.

EUROSTAT (2015). *Demographic Statistics: A Review of Definitions and Methods of Collection in 44 European Countries*. Eurostat.

Lanzieri, G. (2013). On a New Population Definition for Statistical Purposes Note. In *CES Group of Experts on Population and Housing Censuses*. UNECE.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. Brooks/Cole, second edition.

NISRA (2015). Northern Ireland Census 2011 Quality Assurance Report. Technical Report March, The Northern Ireland Statistics and Research Agency.

ONS UK (2017). ONS Census Transformation Programme Annual Assessment of ONS ' s Progress Towards an Administrative Data Census. Technical Report June.

Rao, J. N. K. (2005). *Small Area Estimation*. Wiley, first edition.

Stanghellini, E. and van der Heijden, P. G. M. (2004). A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrica*, 60(510-516).

Statistics New Zealand (2014). Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey. Technical report, Statistics New Zealand.

Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, 81(394):338–346.

Zhang, L.-C. (2019). A Note on Dual System Population Size Estimator. *Journal of Official Statistics*, 35(1):279–283.

Zhang, L.-C. and Dunne, J. (2018). Trimmed Dual System Estimation. In Bohning, D., van der Heijden, P. G. M., and Bunge, J., editors, *Capture-recapture methods for the Social and Medical Sciences*, chapter 17, pages 237–258. CRC press.

Zwane, E. N., van der Pal-de Bruin, K., and van der Heijden, P. G. M. (2004). The Multiple-record Systems Estimator when Registrations refer to Different but Overlapping Populations. *Statistics in Medicine*, 23(14):2267–2281.