UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Social Statistics and Demography

The Irish PECADO project: Population Estimates Compiled from Administrative Data Only

by

John Dunne

Thesis for the degree of Doctor of Philosophy

July 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Social Statistics and Demography

Doctor of Philosophy

THE IRISH PECADO PROJECT: POPULATION ESTIMATES COMPILED FROM ADMINISTRATIVE DATA ONLY

by John Dunne

This thesis proposes a new system of Population Estimates Compiled from Administrative Data Only (PECADO) for Ireland.

Ireland does not have have a Central Population Register (CPR) upon which to develop population estimates in the manner of the Scandinavian and Dutch models. Ireland does have a strong system of Person Identification Numbers (PIN) that are used across pubic administration systems when a person interacts with public services. To the knowledge of the author no statistical agency in the absence of a Population register has yet compiled population estimates using administrative data only.

The PECADO system of population estimates takes as its starting point the compilation of a Statistical Population Dataset (SPD) from administrative data sources using a signs of life (SoL) approach. The SoL approach only includes persons in the SPD where there is strong evidence that a person is alive and living in the state for a significant part of the reference year. The SPD is compiled with respect to a reference year. The SoL approach does not accept a person's registration on an administrative system as sufficient evidence for including that person in the SPD. The SPD counts are then adjusted for undercoverage using an adaptation of Dual System Estimation (DSE) methods. The second list or list B that is used in the PECADO DSE setup also comes from an administrative data source not previously used in the compilation of the SPD.

The thesis considers the traditional DSE approach in the context of a Census under coverage survey (UCS) (Wolter, 1986) and presents an alternative formulation of DSE methods that allows a relaxing of the strict assumptions associated with the traditional approach. This alternative formulation now facilitates DSE methods being applied in a much broader set of circumstances, in particular, where one list is derived from administrative data sources and the second list acts as the capture list where each person in the population has an equal probability of being caught (*homogeneous capture assumption*). The thesis then proposes an extension to the DSE methods, Trimmed Dual System Estimation (TDSE), that provides a tool to allow for the evaluation of suspect parts of the SPD for erroneous records. The thesis also considers the situation where the *homogeneous capture assumption* is weak and discovers that in certain situations this assumption may also be relaxed without introducing bias to the population estimate. We label these set of tools and methods the PECADO toolkit.

The thesis presents and evaluates a set of population estimates using the PECADO toolkit. In particular, the thesis shows this system of population estimates to be plausible. The plausibility of the estimates is demonstrated by using the tools in the PECADO toolkit to provide reassurance the SPD does not contain erroneous records and to provide some reassurance that the second list compiled from administrative data sources does not introduce bias. The population estimates show some differences when compared with the official Census counts. The conceptual differences can be reconciled when migration is factored in but the differences between the estimates are too large for this to be the sole explanation.

The underlying DSE methodology is also considered with respect to using an administrative data list in place of a second field operation to adjust traditional Census counts for undercoverage. We consider the 2016 Census. Ireland to date has not conducted a Census coverage survey and has relied on the diligence and motivation of the Census field force to ensure everybody is counted. Ireland plans to conduct a traditional UCS with a field operation as part of Census 2021. If an administrative list can be used instead of a second field survey then this approach will save significant time and money while also simplifying the application of underlying DSE methodologies.

The PECADO system of population estimates is then further extended to estimate population flows, reusing the same administrative data sources as were used to compile population (stock) estimates. The resulting estimates of population flows (inflows = births + immigration, outflows = deaths + emigration) readily reconcile with the PECADO population (stock) estimates as this extension incorporates a coherent demographic accounting framework.

The thesis concludes with some consideration of possible next steps that could be taken to disaggregate population estimates by detailed geography, how to incorporate attributes and how to build household units with a view to being able to compile Census like estimates on an annual basis.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, John Dunne , declare that the thesis entitled *The Irish PECADO project: Population Estimates Compiled from Administrative Data Only* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: (Zhang and Dunne, 2018)

Signed:...

Date:..

Acknowledgements

First and most importantly, thanks to my family, Rose, Lananh, Yemma, Sofia and David for your support, time and patience.

Also a big thank you to the CSO for the significant support received to undertake this work. I sincerely hope that CSO will in turn be a big net beneficiary of this work.

To University of Southampton, thank you for hosting me.

To my many colleagues and friends, both within the CSO, University of Southampton and across the broader statistical community, thank you for the many chats, proof readings, discussions, exchange of ideas and words of encouragement.

And last but not least, Lichun, what can I say ...

List of Abbreviations

A	Notation to indicate a list in a DSE setup. Typically used to refer to the *fixed* list when used in the DSE setup as per Zhang and Dunne (2018).
AB	Notation to indicate the list of matches between list A and list B.
ABS	Australian Bureau of Statistics, the NSI for Australia. `http://www.abs.gov.au`
ADC	Administrative Data Centre, the name given to the internal unit at CSO, ireland with responsibility for managing administrative data.
AGR	A comprehensive building and address register compiled as part of Census, 2011 in Germany.
B	Notation to indicate a list in a DSE setup. Typically used to refer to the list that satisfies the *homogeneous capture* assumption when used in the DSE setup as per Zhang and Dunne (2018)
CB	Short hand to denote a data source derived form the Irish Child Benefit public administration system
CSO	Central Statistics Office, Ireland. The name of the NSI in Ireland.
CF	Count Factor. Context is the Census in Spain in 2011.
CI	Confidence Interval
CPR	Central Population Register. CPR is available in some countries as part of their public administration systems. Residents are typically required to register and de-register when they move places in order to receive public services and pay local taxes.
DESTATIS	Federal Statistics Office of Germany `https://www.destatis.de`.

D_i The number of records classified as *doubtful* on the Spanish CPR for group i. Context is methodology underpinning the Census in Spain, 2011 (Argüeso and Vega, 2014).

d_i The estimated number of *doubtful* records from Central Population Register for group $i$found in a coverage survey. Context is methodology underpinning the Census in Spain, 2011 (Argüeso and Vega, 2014).

DLD Driver Licence Dataset. A dataset comprising of persons renewing or applying for a new driving licence in the calendar year. Subscripts are used to distinguish years when necessary.

DSE Dual System Estimation. A statistical methodology commonly used for population size estimation.

DSP Department of Social Protection, Ireland. Recently renamed to Department of Employment Affairs and Social Protection (DEASP) http://www.welfare.ie

ECCE Short hand to denote a data source derived form the Irish Early Childhood Care public administration system

EIRCODE A unique code assigned to each letterbox in Ireland. http:\www.eircode.ie.

ERP Estimated Resident Population. The name given to a population concept used by Statistics New Zealand

FET Short hand to denote a data source derived form the Irish Further Education enrolment system

FSO Federal Statistical Office, Switzerland. The NSI for Switzerland. https://www.bfs.admin.ch

HEA Short hand to denote a data source derived form the Irish Higher Education enrolment system

IAF Improved Administrative File. An SPD created by the Israeli Statistical office by enhancing the CPR with other administrative data sources.

ICF Integrated Census File An enhanced SPD created from the ICF as part of the Israeli Census. The ICF is used in the compilation of population estimates for any group or area.

IDI Integrated Data Infrastructure. name given to the statistical infrastructure/project used at Statistics New Zealand for linking administrative data.

IDI-ERP A variant on the ERP population concept used as part of the IDI project at Statistics New Zealand.

INE Instituto Nacional Estadistica, Spain. The NSI for Spain http://www.ine.es.

IPR Israeli Population Register. Name given to CPR in Israel.

ISS Irish Statistical System.

IT Short hand to denote a data source derived form the Irish Self Employed Tax Returns system

K represents the number of Stayers in SPD in a given year. Subscripts are used to denote context. Context is the development of methodology to estimate population flows.

k represents the number of suspicious records in a chosen suspicious part of List A. The context is TDSE.

k_1 represents the number of k suspicious records found in list B. The context is TDSE.

LFS Labour Force Survey. The title of a new household survey designed to measure unemployment in Ireland from Q3 2017 onwards. LFS is a commonly used survey title for similar surveys around the world.

m use to denote the size of the match between two lists typically list A and B. Subscripts are used in places to denote the lists the match is referring to.

MAR missing at random.

MIS Matching Impact Study. The name given to a project undertaken to evaluated new matching techniques as part of Census 2013 in New Zealand.

MoM Method of Moments

N Refers to the population size of U. various subscripts are used to differentiate between different populations.

n Typically used to refer to the size of List B in the Zhang Dunne DSE setup. Subscripts are used when it is necessary to differentiate between different groups.

NDI National Data Infrastructure. The conceptual framework promoted by the NSB to underpin the development and evolution of public administration systems in Ireland. the concept has at its core the use of permanent official identifiers for persons, businesses and properties when engaging with the State.

NSB National Statistics Board. Name given to the advisory board that oversees the strategic direction of CSO, Ireland. http:\www.nsb.ie

NSI National Statistical Institute. The generic name associated with the organisation that is primarily responsible for the compilation and dissemination of Official Statistics in a State.

OCR Optical Character Recognition - a computer technology that is able to recognise hand written and printed characters.

ONS Office for National Statistics. NSI with responsibility for compilation of official Statistics in UK. https://www.ons.gov.uk

p Generally used to denote the probability a unit in the population belongs to a certain subset of the population. Subscripts are used in distinguishing different groups in the population.

P35 Short hand to denote a data source derived form the Irish Employer Employee Tax Returns system

PECADO Population Estimates Compiled from Administrative Data Only. The name given to the title of the project that the thesis describes.

PAR Person Activity Register. The PAR is the name of the SPD used for the compilation of population statistics in the PECADO project at CSO, Ireland. A subscript may be used to indicate a calendar year. Subscripts 1 and 2 are used in the project to denote consecutive years.

PCRS Primary Care Reimbursement System. A data source derived from the Irish public health administration system in Ireland.

PES Post Enumeration Survey. A survey usually undertaken immediately after the final stages of a Census field collection operation with the purposes of investigating the coverage of a Census.

PIN Person Identification Number. A generic name for person related official identifiers used when a person engages with Public Service.

PPP Short hand to denote a data source derived form the Irish Secondary School enrolment system

PPSN Personal Public Service Number. The official PIN for persons in Irish Public Administration Systems. Each person has their own PPSN (assigned shortly after birth) which they typically used when engaging with Public Service systems.

QNHS Quarterly National Household Survey. This survey was the primary household survey designed to measure unemployment in Ireland up until Q3 2017. It was then replaced by a new survey called the Labour Force Survey.

r represents the number of erroneous records in list A, typically unknown.

RSI Revenue Social Insurance. An identification number for persons before the PPSN came into existence in Ireland.

S_i The number of records classified as *sure* on the Spanish CPR for group i. Context is methodology underpinning the Census in Spain, 2011 (Argüeso and Vega, 2014).

s_i The estimated number of *sure* records from Central Population Register for group $i$found in a coverage survey. Context is methodology underpinning the Census in Spain, 2011 (Argüeso and Vega, 2014).

SoL Signs of Life. Typically refers to evidence that somebody is alive based on an interaction with a public administration system. The concept can be extended to include a broader set of signals.

SNZ Statistics New Zealand, the NSI for New Zealand `https://www.stats.govt.nz/`.

SP Short hand to denote a data source derived form the Irish State Pension system

SW Short hand to denote a data source derived form the Irish Social Welfare system

TDSE Trimmed Dual System Estimation, an extension of DSE methods that enables the hunting for erroneous records in List A in the Zhang and Dunne DSE setup. Zhang and Dunne (2018)

U Refers to the reference population. Various subscripts are used throughout the project to denote different populations.

UCS Undercoverage Survey. General term for a survey designed to measure undercount in a Census. Can also refer to undercount in an SPD.

UNECE United Nations Economic Commission of Europe http://www.unece.org

UR Usually Resident. used in the context of a qualifier for the population concept, e.g, usually resident population includes persons having resided or intending to reside for a period of 12 months.

WCF Weighted Census File. the name used for the final SPD in the Spanish 2011 Census where each record has a weight attached such that summing over the weights for particular groups will provide an estimate of the population size for that group (Argüeso and Vega, 2014).

x Refers to the size of a list (or part of a list) in a DSE setup. In the context of the DSE setup used by Zhang and Dunne (2018), x refers to the size of list A only, the *fixed* list.

δ A 0,1 indicator variable for each unit in the population to denote whether that unit is also included in a list. A subscript i taking values 1 to N is used to denote the relevant unit in the population and a second subscript may be used to denote the specific list the indicator refers to.

π Probability of a unit of the population being included in a list. Can be unknown. The notation is used in the Zhang Dunne DSE setup to denote the constant capture rate (*homogeneous capture*) with List B for each unit in the population. Subscripts can be used to denote different population groupings.

ξ_x the ratio of those persons present in the SPD in two consecutive years over the total number of persons that appear in the SPD over two consecutive years. Context is the development of methodology to estimate population flows.

ξ_N the ratio of those persons present in the population in two consecutive years over the total number of persons that appear in the population over two consecutive years. Context is the development of methodology to estimate population flows.

θ	represents the odds ratio. Used in a number of contexts in this project.

Chapter 1

Introduction

1.1 Why a new system of population estimates?

The Central Statistics Office, Ireland (CSO) enumerated 4.76 million people living in the Republic of Ireland in 2016. The associated Census cost in excess of €60m, or over €12 for every person living in the State.

For countries that do not have a Central Population Register (CPR) from which demographic statistics can be compiled, 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 typical approach to population estimates in these scenarios is an application of the demographic component or cohort component method to postcensal population estimates and then a recalibration of population estimates for intercensal estimates. In a Eurostat review of 31 countries published in 2003 (EUROSTAT, 2003), 19 countries were identified as using the component method for population estimates. This approach 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. This approach is typically applied to each of the different 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

	thousands				
Year	2011 to 2012	2012 to 2013	2013 to 2014	2014 to 2015	2015 to 2016
Component					
Population at timepoint 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 timepoint 2	4,593.7	4,614.7	4,645.4	4,687.8	4,739.6

Table 1.1: Population estimates and their components for Ireland (thousands). Source: Central Statistics Office, Ireland (http://www.cso.ie).

intervals. The Census provides a benchmark to recalibrate the population estimates at regular intervals. 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. Table 1.1 provides an overview of the estimated population and change components for Ireland over the years 2011 to 2016. These estimates are intercensal estimates compiled after Census 2016 was completed.

In Ireland, the principal source of information for the estimation of the gross annual migration flows was the Quarterly National Household Survey (QNHS), which also provides the basis for the classification of the flows by sex, age group, origin/destination and nationality. The QNHS targeted 25,000 households (somewhere between 1% and 2% of households in the state) each quarter. The QNHS was replaced by a new quarterly Labour Force Survey (LFS) in Q3 2017 with a similar sample design. The migration estimates are also compiled with reference to movements in other migration indicators such as the number of Personal Public Service Numbers (PPSNs) allocated to non-Irish nationals and the number of visas issued to Irish nationals with respect to a number of destinations including Australia, US and Canada. In addition, data on National Insurance numbers (UK equivalent to PPSNs for tax purposes) issued to Irish nationals in the UK is considered.

Given the relative size of migration flows at $< 3\%$ (see table 1.1) and the QNHS sample size ($< 2\%$) there are considerable challenges with estimating migration flows. Given that these migration flows currently contribute to the compilation of population estimates any new contributions that can enhance the quality of these estimates will have significant value in their own right.

Therefore, if it is possible to compile reliable population estimates on an annual basis then this would negate the requirement of conducting a Census every 5 years. Ireland could 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.

This thesis proposes a system of Population Estimates Compiled from Administrative Data Only (PECADO). It is novel in that to the knowledge of the author no country has yet compiled official population estimates solely from administrative data sources where no Central Population Register (CPR) exists. The thesis then considers the use of the underlying methodologies in the proposed PECADO system to explore undercoverage within the traditional Census setting in a timely and low cost manner. To date, the Irish Census assumes undercoverage is not an issue in enumerating the population. However, CSO, Ireland plan to incorporate an undercoverage survey (UCS) as part of the Census in 2021. The thesis also considers an extension of the proposed PECADO system to estimate population flows (and hence migration), again using administrative data only.

1.2 National and International context

1.2.1 International Context

1.2.1.1 Overview

For many countries, the Census is the backbone in their system of population estimates.

For the 2000 round of Censuses only 4 of 44 countries conducted a Census where the enumeration was based solely on registers, as recorded in the 2008 United Nations Economic Commission of Europe (UNECE) survey (UNECE, 2008). For the other 40 countries that conducted the Census in the traditional manner, the Census was considered an integral part of the National Statistical System. The traditional Census was integral in that it provided and updated sampling frames and statistical registers, along with providing a considerable amount of information on each household and person in the country at a particular point in time. For many countries it is also the only source of reliable small area statistics and particular subject matter domains (including those with relevance for small sub-populations).

Conducting a traditional Census is a major logistical exercise presenting many challenges. These challenges include cost, organisation and timeliness. Many countries looked to the experience of the 4 countries that did not have to contact each household and noted the potential to mitigate these challenges.

A substantial element of the cost of a traditional Census is attributed to employing a field force to ensure each household in the State is enumerated. In a register based Census, persons and households are simply counted using the CPR. The CPR, available in some national administrations, is the backbone to the co-ordination and delivery of public services to individuals and households.

Given the costs associated with a traditional Census, most countries will only conduct a Census every 10 years. Some countries such as New Zealand and Ireland conduct a Census every 5 years. Scaling up to conduct a traditional Census in this manner provides considerable organisational challenges that are disruptive to the annual planning cycle of National Statistical Institutes (NSIs). NSIs have to secure funding, re-organise budgets, recruit staff and reallocate other experienced statistical and technical resources to ensure that the Census is conducted in an effective manner. Technology and systems used for a previous Census are typically obsolete and need to be replaced or significantly upgraded. Therefore, there are significant organisational and resource benefits to negating the requirement for a traditional Census. Cost and increased efficiency were the key drivers behind Denmark's move away from a traditional Census (Lange, 2014).

Other drivers that motivated the first countries to move away from a traditional Census include the difficulty involved in contacting each household and the burden and intrusiveness associated with enumerating each person in the State. In 1971, the Netherlands experienced significant privacy objections with the intrusive nature of the Census and this along with the significant cost savings (estimated at €3m compared with €300m for a traditional Census) motivated the move to a *virtual Census* (Nordholt, 2005).

Another significant benefit to moving to a lower cost register based Census is the ability to be able to produce Census type population statistics on an annual basis (UNECE, 2007). However, there are some drawbacks to this modernisation. These drawbacks include a reliance on the information content and structure available through the registers, difficulties in mapping administrative concepts to statistical concepts, timeliness issues with respect to the availability of registers and, finally, no longer having the capacity to ask new questions of every household with respect to emerging statistical needs.

For the 2010 round of Censuses, of the 54 countries that participated in the UNECE survey, 34 countries were identified as conducting a traditional Census, while 19 countries were identified as conducting a register based or combined Census, where a combined Census is categorised as data from registers combined with a field collection (UNECE, 2014). France introduced a rolling Census. The use of registers and administrative data had increased significantly. Furthermore, as a follow up to the 2010 round of Censuses, the UNECE survey identified 15 countries with a traditional Census that will include registers as part of the methodological design for 2020, with 13 of those countries stating they will collect Census data from administrative sources.

Population flows are also an important consideration for any system of population estimates. Population flows can be broken into two components; population *inflows*, those persons joining a population from one timepoint to the next and population *outflows*, those persons leaving a population from one timepoint to the next. Population inflows generally comprise births and immigration in the reference period whereas population outflows comprise deaths and emigration in the same reference period. This overall relationship underpins the component method described earlier in section 1.1.

When population estimates are compiled directly from administrative registers, the associated estimates of population flows are coherent. The flows are simply identified and counted by comparing the population register at two points in time. Accurately identifying deaths and births also enables coherent estimates of migration flows (emigration and immigration).

In the absence of suitable administrative registers, the compilation of reliable population estimates relies on being able to estimate population flows properly. For most countries, births and deaths are typically registered and easily counted. Therefore, it becomes important to be able to compile reliable migration estimates. For those countries with strong border controls and recording systems the compilation of such migration estimates is theoretically easier, but not without its problems. Israel is one such country where data recorded at border controls is used in the compilation of population statistics (Central Bureau of Statistics of Israel, 2015).

In estimating migration flows, immigration is typically easier to estimate than emigration. Immigrants are resident in the country and can be picked up and estimated through various surveys and administrative data sources. Emigration is more difficult to estimate as emigrants are no longer present in the country and are not picked up in surveys or administrative sources. In his review of methods for estimating emigration, Jensen (2013) considers different approaches including register based, residual based, survey based, indirect estimation and modelling.

For the reasons above, countries that cannot compile population estimates using registers need to rely on a traditional Census to provide reliable population estimates at frequent intervals. These Census population estimates, along with estimates of migration, then form the basis of intercensal and postcensal population estimates.

In practice the Census also encounters its own challenges in enumerating every person in the population. Most countries accept that the traditional Census can suffer from coverage problems and as such conduct a Post Enumeration Survey (PES), often referred to as a coverage survey, to evaluate and adjust for coverage issues (UNECE, 2008, 2014). Typically, the traditional Census will suffer from net undercoverage. For example, Statistics New Zealand estimate they counted 96.7% of the persons usually resident in New Zealand on Census night, 2013 (Statistics New Zealand, 2014a). In comparison, register based population estimates can tend to suffer from overcoverage issues where emigrants may fail to deregister from the underlying population register. The experiences of Israel in developing a 2008 register based Census (Kamen, 2005) illustrate this point, where some 7% of those registered in the population registry were judged to have emigrated.

The UNECE surveys (UNECE, 2008, 2014) categorise countries into 3 groups, register based, combined and traditional, with France being an exception to this categorisation having implemented a rolling Census (Durr, 2005). Taking the trend of *Census modernisation* as having, or developing, the capability to conduct a Census using administrative data to mitigate the requirement to directly enumerate every person in the country, we will now look at the system of population estimates for a selection of countries categorised under 3 headings as follows.

- Census modernisation - mature
  Those countries that have conducted their Census enumerations directly from administrative sources or registers for a significant number of years. Examples of these countries are the Nordic countries and the Netherlands.

- Census modernisation - first steps taken
  Those countries that have already taken the first steps in compiling population estimates without trying to contact every household to enumerate every person in the State. Administrative data sources play a significant role in the compilation of population estimates. Example countries include Israel, Spain, Germany and Switzerland.

- Census modernisation - aspiring
  Those countries that recognise the potential of administrative data sources and are actively investigating systems and methods to conduct a Census for the first time without trying to contact every household to enumerate every person. Example countries include UK, New Zealand and Estonia.

1.2.1.2 Census modernisation - mature: Nordic countries and Netherlands

Nordic countries: Denmark, Finland, Sweden and Norway

The register based system of statistics in the Nordic countries can be traced back to concepts originally developed by Nordbotten (Nordbotten, 2010) in the 60's and was probably best described by the simple model for a socio-demographic statistical system presented by Thygesen (Thygesen, 2010) in the early 80's. See figure 1.1. In fact, this

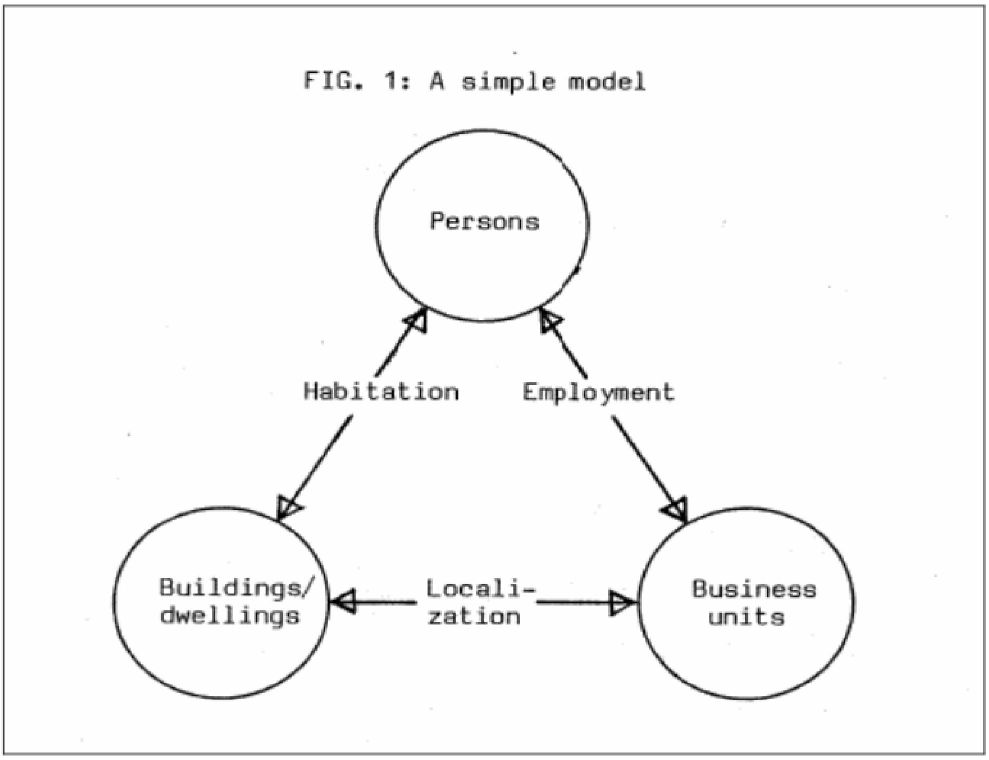

Figure 1.1: Representation of socio-demographic statistical system presented by Thygesen in early 80s (Thygesen, 2010)

representation and variations of it are increasingly being used to explain how to organise data in a register based statistical system with official identification numbers for each of the different types of statistical unit (Persons, Building/Dwellings and Businesses).

The Nordic register based statistical system becomes possible as the municipalities actively use population registers in the delivery of public services. By law, residents are required to register and deregister with government municipalities as they move in and out of them. The law is further reinforced through the use of record extracts from the register as evidence in the conduct of different administrative activities, such as applying for a passport, getting married or divorced. This system is managed through the use of official identification numbers for both persons and buildings or addresses.

The Nordic system is well documented with a report compiled by experts providing a review of best practices in compiling population and social statistics in 2007 (UN-ECE, 2007). The register based system in the Nordic countries did not come into place overnight and was developed on a step by step basis over a number of years. Thygesen (2010) and Lange (2014) provide an interesting history and understanding of how this system developed in Denmark, including the many difficulties that were surmounted along the way. In particular, Thygesen (2010) recalls the internal discussions in Statistics Denmark concerning the paradigm shift away from the traditional Census, recalling a quote from one of the pioneers of this new system 'On what grounds can anyone claim that there has ever been one single piece of correct information recorded on a Census form?' from the 1979 Nordic statisticians meeting. A significant milestone for Statistics Denmark was when Eurostat supported the translation of and published the book 'Statistics on persons in Denmark. A register-based statistical system'. Today, according to claims made by Lange (2014), the Census results are compiled by only two persons. This is predicated on already having the data collected and properly organised as required in a fully functioning register based statistical system.

Nordbotten (2010) provides a broader history of the underlying ideas and also discusses some of the privacy issues. Nordbotten also notes a 1960s proposal for a national archive center in the US and notes the heated privacy debates that ensued. Kraus documents this proposal and the privacy concerns raised, with a view to informing future policy decisions at the US Census Bureau (Kraus, 2013). The privacy concerns raised then are still relevant today.

Statistics Sweden recognises that one weakness of the CPR is that it may not record deregistrations from emigrants in a timely or accurate manner. Recent work (Bengtsson and Rönning, 2016) uses the concept of *imprints*, or as we refer to later *Signs of Life (SoL)*, in administrative data sources to explore overcoverage issues and reassuringly demonstrates that potential overcoverage is significantly less than 1% of the population in the case of Sweden. Their work considers a number of different indicators based on activity for a person in the 2 years before and after the reference year. A person identified with no activity identifies suspected overcoverage. It does not equate to a person not belonging to the population but certain indicators will indicate a higher likelihood of belonging to an overcoverage group. For example, a person with no activity but who has graduated from a third level course two years previously will be given a high likelihood of having emigrated. Records are then weighted based on the indicators. This weighting approach then shrinks the number of persons without an imprint to a more plausible estimate of the population.

Netherlands

As documented by Nordholt (2005), Statistics Netherlands conducted their last traditional Census in 1970 where they experienced considerable privacy difficulties. One of the roles of the old Dutch Census was to update the municipal registers but as the quality of the registers increased the role of the Census diminished in this regard. Subsequent Censuses were conducted using registers and existing surveys. It is generally accepted that it is very difficult to live in the Netherlands without being registered on the CPR. Municipalities are also motivated to keep the CPR up to date as it underpins the allocation of funding from Central Government. These two factors combine to ensure registers are of a high quality. The CPR underpins the official population of the Netherlands.

Statistics Netherlands has also considered coverage problems when moving from an official population definition concept (as defined by the CPR) to the statistical definition of usual resident population (as defined in Regulation 1260/2013) (Statistics Netherlands, 2016). Two solutions to addressing coverage issues are considered and a combination of both solutions is discussed. The first solution, described in more detail by Gerritse et al. (2016), considers an application of capture-recapture methodologies to explore and address undercoverage issues. The second solution, termed *micro register data method*, simply adds records to or removes records from the CPR to obtain the usual resident population. No mathematical estimation is undertaken in this second approach. The rules for adding or deleting records are based on activity of persons in other administrative registers.

Both approaches have their advantages and disadvantages. The first solution, based on capture-recapture, provides estimates of undercoverage but does not provide any estimate of overcoverage. This first solution only provides an estimate for undercoverage at the national level. The second solution, while providing estimates of undercoverage and overcoverage, will underestimate both. The second solution fails to pick up information about a number of groups such as illegal or undocumented workers.

Statistics Netherlands uses a combination of the two solutions to estimate undercoverage. Undercoverage is estimated at the national level using capture-recapture methodologies and the micro register data method is then used to disaggregate the undercoverage estimate by region. In summary, Statistics Netherlands (Statistics Netherlands, 2016) conclude that the usual resident population was approximately 16.9 million in January 2013, 0.8% higher than the official population. Undercoverage and overcoverage in the official population is estimated at 169,900 (1.0%) and 33,200 (0.2%), respectively, when considering the estimated usual resident population.

For a number of reasons, Statistics Netherlands cautions that neither the old (based on official population estimate) nor new usual resident population estimate can be considered definitive. The reasons include uncertainty with one of the underlying data sources (the Crime Suspect Register) used in the capture-recapture method, missing data associated with some groups and the assumptions that need to be made when applying either method. Statistics Netherlands also states that they reserve the right to revise these estimates should new methods or data become available. This approach is also considered time consuming and, at the moment, Statistics Netherlands does not consider it practical to undertake on an annual basis. They do, however, suggest this approach should be repeated after a number of years for validation purposes.

1.2.1.3 Census modernisation - first steps taken: Israel, Spain, Germany and Switzerland

Israel

Israel took first steps in 2008 to modernise the Israeli Census (Kamen, 2005) and are now preparing for their second register based Census in 2020 (Blum and Feinstein, 2017). The Israeli Census is referred to as an integrated Census and in summary consists of combining the Israeli Population Register (IPR) with a 20% sample of households to collect attributes in tandem with estimating undercoverage and overcoverage.

The IPR contains a record with an official identification number for all persons officially resident in Israel, past and present. The quality of the address information in the IPR is considered unreliable (25% of persons on the IPR are estimated to live at a different address than the residential address recorded on the IPR). The IPR is updated with information from other administrative data sources to create an enhanced Statistical Population Dataset (SPD) called the Improved Administrative File (IAF). The IAF now forms the population spine from which population estimates are compiled.

For the 2008 Census, undercoverage and overcoverage were estimated through the use of a representative sample comprising 20% of Enumeration Areas (EAs), small geographically contiguous areas comprising of approximately 50 households, and comparing the list of persons identified at each address within an EA with the list of persons in the IAF registered as living at that EA. Undercoverage in the IAF for the EA is then generally estimated by identifying persons living in the selected EA but having an address recorded as elsewhere in the IAF. Overcoverage in the IAF for the selected EA is then estimated by identifying persons that are recorded as living in the identified area but who are not found living there.

The estimates of undercoverage and overcoverage from the selected EAs are now used to estimate undercoverage and overcoverage rates in similar EAs in the rest of the country. Stratification or grouping of homogeneous EAs is used for this purpose. Each record in the IAF is then assigned a weight based on the recorded address and estimates of undercoverage and overcoverage rates and stored in a new file called the Integrated Census File (ICF). The population for any group or area is then estimated by summing the weights for that group or area in the ICF. The weights can be adjusted depending on considerations of the population estimates when compared with alternative estimates.

The following conditions underpin the population estimates for each statistical area (Kamen, 2005). Statistical areas are the smallest geographical area for which population estimates are compiled and comprise of one or more EA.

- No erroneous enumeration in the survey
- Independence between the IAF and survey
- All persons have equal probability of being listed in the IAF, within EA group or stratum

- All persons have equal probability of being enumerated in the field, within EA group or stratum
- Distribution of overcoverage of any statistical area across its EAs is proportional to the distribution of the *true* population

This dual list method in 2008 depended on formal address systems which only covered about 70% of the population. Other solutions are required to enumerate the remaining 30% of the population. There were also difficulties with estimating some sub-groups such as immigrant workers, undocumented residents, or workers and foreign students.

The reference population is all persons with an official identification number excluding those that had been abroad for a year or more plus those persons without a reference number who had been present in Israel for at least a year. The latter group are estimated by simply adding in any person found in field operations to the ICF and assigning them a weight of 1.

To produce annual population estimates, a new IAF was created and weights w for each estimation group i in the IAF are updated. For example, weights for the 2016 population are calculated as follows:

$$\hat{w}_i^{2016} = \frac{P_i^{2008} + \hat{C}_i^{2008-2016}}{IAF_i^{2016}}$$

where

P_i^{2008} is the population in the Census in 2008,

$\hat{C}_i^{2008-2016}$ is the changes in the CPR from the last Census until the end of the reference year and

IAF_i^{2016} is the IAF count at the end of 2016.

Israeli plans for the 2020 Census (Blum and Feinstein, 2017) include updating the methodology with regard to assigning weights to individual records, use of new administrative data sources and enhanced estimates of the foreign population. Israel is also planning to include *SoL* type data for improving the local coverage estimation in the Census.

Spain

Spain's municipality registers were set up in 1996 to record all persons resident in each municipality regardless of legal status. The registers are also used to provide official population figures at the municipality level.

In 2010, the Instituto Nacional de Estadistica, Spain, (INE) conducted a population and housing Census based on registers and a 10% sample survey (Argüeso and Vega, 2014). The sample survey was designed to estimate coverage errors as well as collect attribute information on the population. A building Census enabled geo-referencing of each building.

The base register was compiled by integrating each of the population registers from the different municipalities. There were approximately 47.3m persons in the base register, of which approximately 5.3m were recorded as foreign nationals. This was then enhanced by examining different administrative data sources (tax, social security, vital events, etc.) to create an indicator of proof of residence to be included on the Census file. 2.2% of persons from the base register were identified as having no proof of residence from other sources and were classified as *doubtful*, 0.1% were identified as being erroneous or deceased and were excluded, while the remaining 97.7% of persons were classified as being *sure*. 87% of the approximately 1m doubtful records were recorded as having a foreign nationality. The survey was then used to assign weights or count factors to the 2.2% of doubtful records on whether they were part of the population or not and included in a Weighted Census File (WCF) where sure records have a weight of 1. Population estimates for any group are now compiled by simply summing weights attached to each record in that group.

To calculate the weights to be associated with the *doubtful* records, the following steps were taken. The survey data was first classified by age, nationality and geography into a number of groups. The survey data was then linked with the base register data at an individual level to identify which records in the survey to classify as *sure* records. The remaining records in the survey data are classified as doubtful. A weight or count factor (CF) was then calculated for each doubtful record based on the group it belonged to and included against that record in the WCF. The count factor CF_i where i denotes the group was calculated as follows:

$$CF_i = \frac{d_i}{s_i}\frac{S_i}{D_i} \tag{1.1}$$

where

S_i is the number of sure records in the base register for group i

D_i is the number of doubtful records in the register for group i

s_i is the estimated number of sure records in the sample for group i

d_i is the estimated number of doubtful records in the sample for group i

ensuring the total number of persons in each group i could be estimated by $\hat{T}_i$ as

$$\hat{T}_i = S_i + CF_iD_i \tag{1.2}$$

It is possible that a weight for a group is greater than 1. This can happen if the register had undercoverage that resulted in the survey sample including persons that were not registered. In fact, if we substitute equation 1.1 into equation 1.2, it can now be written in the form of a Dual System Estimator (DSE) in equation equation 1.3 as follows:

$$\begin{aligned}\hat{T}_i &= S_i + CF_iD_i \\ &= S_i + \frac{d_i}{s_i}\frac{S_i}{D_i}D_i \\ &= S_i\left(1 + \frac{d_i}{s_i}\right) \\ &= \frac{S_i(s_i + d_i)}{s_i}\end{aligned} \tag{1.3}$$

where

S_i corresponds to list A with only undercoverage.

$s_i + d_i$ corresponds to list B, a sample from the population.

s_i corresponds to the match between list A and list B.

This approach to the Census is estimated to have cost approximately €85m resulting in significant savings. A traditional Census would have cost in the region of €500m to €550m .

The population was estimated at 46.8m or 450,000 less than the registered population as at 1st November 2011. This difference is less than 1%.

Annual population estimates are now compiled using the component method with the 2011 Census figures used as the starting point (INE Spain, 2014) and are published with reference to the 1st of January and the 1st of July each year. INE Spain have also considered statistical solutions to the problems with de-registration of emigrants in compiling estimates of emigration (INE Spain, 2018).

Germany

In 2011, the Federal Statistics Office of Germany (DESTATIS) conducted a register based Census and used a survey to correct for undercoverage and overcoverage and assure the quality of results (Bechtold, 2016). The official population was enumerated as approximately 80.2 million persons on May 9, 2011. The register based Census was conducted without the use of personal identification numbers or building identification numbers. All linking was done on the basis of matching records on name, address, date of birth, place of birth and other personal characteristics.

The key drivers to conducting a register based Census in 2011 were cost and response burden. The experiences of conducting Censuses in the 1980s in West Germany, where privacy debates and concerns led to the boycott of the 1983 Census and the postponement of the 1987 Census, were a significant consideration in the decision to conduct a register based Census in 2011 (Scholz and Kreyenfeld, 2016). Germany did not conduct a Census in the 2000 round. For non Census years, Germany conducts a *microcensus*, a representative 1% sample survey of the population, every year since 1957 (Schwarz, 2001). The first microcensus was undertaken in 1957. The microcensus is a flexible, multi-purpose survey, including Labour Force Survey requirements, that is used to determine population structures between two Censuses.

The compilation of a comprehensive building and address register was considered critical to enabling the linking and geo-referencing of data. The register, called the AGR, was compiled from various administrative data sources including the population registers maintained by the administrative authorities. The linking of persons in the combined population register provided an initial SPD. A survey of households using the AGR as a sampling frame was then used to estimate undercoverage and overcoverage. The AGR is assumed to have no undercoverage with respect to buildings and addresses. The sample survey covered nearly 10% of the population.

The nature of national legislation governing registers constrained the statistical use of these registers in that only a *temporary central population register* could be constructed for statistical purposes. This prevented any post validation of results or methodology development once the Census was complete. Furthermore, the Census was a complex procedure requiring coordination between the different survey components and data sources. Consolidation and linking of data sources proved challenging due to the number of data sources that needed to be linked and the lack of standard identifiers. Bechtold (2016) also acknowledges that the interpretation of the results may be more complex than a Census conducted by a complete enumeration.

Scholz and Kreyenfeld (2016) discuss the Census from a demographic research point of view and attempt to assess the accuracy of the results before considering systematic sources of error in the updated population estimates. In their conclusion they note:

- the timeliness of the final age by sex results from the German Census (4 years) and compare it to the Scandinavian countries where final results are published within one year
- the complex procedures associated with the German Census would question the argument that a register based Census is less expensive and more effective than a Census conducted by traditional enumeration
- the opportunity to use the Census to build and maintain some type of household register and population register over the long term has been missed.

Switzerland

Switzerland moved to register based Census in 2010 (Schwyn and Kauthen, 2009). The move was a nationally co-ordinated one across the different public authorities with new *Census based* legislation passed in 2007 requiring relevant register keepers to incorporate a 13 digit official person identification number on relevant registers and data sources. The system also includes official identifiers for buildings and dwellings making data linking easy in a register based system.

The Census in 2010 was the start of building a new comprehensive system of household and person statistics at the Swiss Federal Statistical Office (FSO). The system can be briefly described as follows:

The register is first collected and collated from the different public authorities each year. This provides the spine for the SPD with all address information geo-referenced to a high quality. An annual structural survey of approximately 200,000 persons is also conducted to collect attributes not available in the register. A suite of topic based surveys are also undertaken each year on a rotational basis. FSO also undertakes an omnibus survey of approximately 3,000 persons. Eichenberger et al. (2010) provide additional information on the anticipated accuracy of the Swiss Population Survey.

In 2013, the FSO undertook a coverage survey to evaluate the coverage errors for persons and buildings in the 2012 Census (FSO, 2015). The coverage survey involved trained personnel visiting selected zones and identifying every building in that zone. The interviewers also interviewed 21,000 households in a survey process that involved a significant degree of promotion and follow up to nonresponse. Capture-recapture methods, also known as DSE methods, were used to evaluate undercoverage.

Undercoverage of buildings in the 2012 Swiss Census was estimated at 0.18% while overcoverage was estimated at 0.71%. Undercoverage of persons was estimated at 0.47% while over overcoverage was estimated at just 0.02%. These figures compare well with the 2000 Census coverage survey results where net under coverage was estimated at 1.4%.

The FSO conducts an annual register based Census. The FSO does not envisage conducting regular quality surveys of their Census. At the end of 2017, the population of Switzerland is estimated at 8.5 million persons.

1.2.1.4 Census modernisation - aspiring: New Zealand, UK and Estonia

New Zealand

Statistics New Zealand (SNZ) conducted a traditional Census in 2013 and counted 4.24 million people on Census night. The Census was postponed from 2011 due to an earthquake. The Census budget was approximately NZ$90m (Statistics New Zealand, 2012). New Zealand typically conducts a Census every 5 years. The rising cost of the traditional Census has driven discussions about the sustainability of the traditional model. SNZ are actively looking at ways to reduce this cost. Other motivating factors include the increasing difficulties associated with contacting everybody and complexities in addressing coverage issues.

SNZ, like many other countries, conducts a PES, to address any coverage issues in the Census. Following the 2013 PES, the number of New Zealand residents present on Census night was estimated to be closer to 4.35 million, a net undercount of 103,800 or 2.4% (Statistics New Zealand, 2014a).

The PES methodology used derives weights for households and persons which are then used to estimate undercoverage and overcoverage for the different population groups. The population groups are described by sex, age group, ethnicity and geography. The sample size for the PES in 2013 was 15,000 households up from 11,000 in 2006. A major innovation to the 2013 PES was the introduction of *automated matching*. In 2006 this was done manually. The matching was done using information on date of birth, sex, ethnicity, names and usual residence address.

SNZ conducted a Matching Impact Study (MIS) to evaluate the new matching methodology. The study involved undertaking a matching exercise using the old manual methodology on a sub sample of the PES and compiling population estimates to compare with the estimates derived using the new methodology. The study suggests that, if the old matching methodology was used, the net undercount rate would be estimated at 3.9% compared to an estimate of 2.4% estimated with *automatic matching*.

The PES methodology also contains a number of other assumptions that cannot be validated within the survey. These assumptions include non response in the PES being considered missing at random and no dependence between a dwelling being missed in the Census and the same dwelling being missed in the PES.

The system of annual estimated resident population (ERP) estimates for New Zealand are based on the component method where migration and natural increase in the population is used to move the population estimates forward a year (Statistics New Zealand, 2014b).

SNZ is actively investigating how to transform the Census. The strategy looks at how to make the current traditional model more efficient while at the same time exploring alternative ways of producing small area population statistics. The strategy includes consideration of moving to a 10 year Census and a Census based on administrative data (Statistics New Zealand, 2012).

SNZ has been working with integrating different administrative data sources for a number of years for research purposes. The project, called the Integrated Data Infrastructure (IDI), has delivered statistical infrastructure of the same name. More recently, SNZ has used the IDI as a test environment to compile and publish experimental population estimates (Statistics New Zealand, 2016). New Zealand does not have a system of official identifiers, and data linkage is undertaken using demographic identification information such as name, address and date of birth to create a population spine. O'Sullivan (2015) discusses and compares the data linking procedures in the IDI, PES and also at Australian Bureau of Statistics (ABS) in detail.

To produce the Estimated Resident Population from the IDI (IDI-ERP), the population spine is used to form a base and a number of filter based rules are then applied to reduce the number of records down to what can be considered the resident population. The IDI spine without any rules applied contains over 9 million persons while the Estimated Resident Population (ERP) is 4.7m for 2017. Inclusion rules are of the form retain any person with an indicator of activity (tax returns, pharmaceutical prescriptions, school enrolment, etc.), while exclusion rules are of the form remove any person who has left the population (deaths, emigrants, etc.).

SNZ also publish the quality targets for the IDI-ERP in the context of the true population (McNally and Bycroft, 2015). This provides a target for the IDI-ERP if using the ERP as a measure of the true population. The IDI-ERP shows potential. However, there is recognition that administrative data and rules alone will not be sufficient and Statistics New Zealand are continuing to progress work developing a coverage survey and statistical models to adjust for errors and discrepancies Statistics New Zealand (2016); Dunne and Graham (2019).

United Kingdom

The Office for National Statistics (ONS), along with its counterparts in Scotland and Northern Ireland, typically undertakes a Census every 10 years for the United Kingdom using a traditional model. The cost of the 2011 Census was estimated at £480m. The Census estimated the population of the United Kingdom to be 63.2 million people.

The United Kingdom also undertakes a significant exercise to estimate and adjust for coverage errors in the Census (Abbott, 2009).

As part of their Census transformation program (ONS UK, 2017), a continuation of the Beyond 2011 program, the ONS has invested heavily in looking at next generation Census models for the Census post 2021. They, like New Zealand, have primarily focused on a rules based approach in developing an SPD that can be used to estimate the size of the population. In their 2017 annual assessment on progress (ONS UK, 2017), ONS highlight the need to use a new legal framework, the Digital Economy Act 2017, to access more activity data to use in combination with a coverage survey to improve the population estimates.

The UK does not have an official identification number that can be used across administrative data sources for linking. The ONS rely on personal information such as name, postcode, date of birth and gender. The linking process is documented in a methodology report (ONS UK, 2013). An added complication to the linking procedures is a requirement to use anonymised match keys.

Estonia

Estonia's pursuit of census modernisation since 2002, the year the Estonian population register was created, is of interest to many countries and is told in a concise an informative way by Beltadze (2020). The experience of Estonia's Nordic neighbours across the Baltic sea has inspired Estonia in its pursuit of Census modernisation. From an initiative launched in 2007, where Estonia analysed the population and housing indicators in 11 State databases, they found that the underlying databases did not have sufficient information to meet the mandatory requirements for the 2011 Census under EU legislation. They also found the databases did not have the necessary additional variables required by Census users. There were many shortcomings in registers and other administrative sources if these sources were to meet the expectations of the 2011 Census.

In 2011, Estonia conducted a *combined* Census whereby each person in the population would be directly enumerated in an operation supported by administrative registers (Tiit, 2014). Administrative registers played a key role in the planning and operational stages of the Census as well as supplementing for missing data and assessing coverage. In assessing coverage, the Census enumerated 1,294,455 persons and with a population size of 1,320,000 estimated by combining registered population events with enumerated persons, resulting in a net undercoverage rate of approximately 2% (Maasing and Tiit, 2019). The Estonian population register had 1,365,000 entries or an overcoverage rate between 3% and 4% when compared to the population estimate. Statistics Estonia published detailed Census tables for the enumerated population in 2011 rather than the estimated population and then for official population estimates published estimates that had been adjusted for undercoverage. One unique feature of the Census 2011 in Estonia was that 68% of the population enumerated was done online, far greater than any other countries in that Census round (UNECE, 2014).

Subsequently, Estonia used the 2011 Census dataset to evaluate administrative registers and to plan for a register based Census in the 2020 round of Censuses. While there were a number of strands to this work, the most interesting is the development of a methodology based on calculating indexes (taking values between 0 and 1) that can then subsequently have a threshold applied, based on reliable training data, that will result in a decision on whether to include a person from administrative registers as part of the population (Beltadze, 2020; Maasing and Tiit, 2019). In principle the index comprises of two components. The first component is best described as a *generalised sum of signs of life* where each person receives a score for each administrative data source they show a sign of life on and then a weighted sum is taken. The second component is a residency index based on information about a persons residency in the previous year. This second component is included to reduce any instability that might occur due to an over dependence on the signs of life approach. A similar index approach has also been applied and assessed with respect to *partnerships* and *placement*(geography).

Following a review in 2019 (Beltadze, 2020), the strategic risk that the results of the register based census will not meet the needs of users in terms of place of residence, also having implications for household and family characteristics, is not an acceptable risk and as such the register base census will be deferred to a later Census round. In summary, data users in Estonia are not yet ready to accept a register based census. However, the prospect of high online response rates married with high quality administrative data sources may provide for significant census modernisation opportunities in the future.

1.2.1.5 Preconditions, challenges and emerging trends with Census modernisation

Traditionally, the key features of a Census are considered to be individual enumeration, simultaneity, universality, defined periodicity and small area statistics. However, with the introduction and consideration of new approaches and broader demands the traditional concept of a Census is being challenged. There is now a demand for more frequent and relevant data at a small area level than is delivered by the traditional Census conducted every 5 or 10 years (UNECE, 2015, 2006). Census modernisation projects are looking at how to meet this demand from administrative data sources.

The long term vision for Census modernisation in many countries is to be able to undertake a Census more frequently (annually), at a much lower cost and with much lower response burden. The inspiration comes from what has been achieved in the Nordic countries (UNECE, 2007). Tønder (2008), based on the UNECE report, considers the

Nordic experience and lists a number of preconditions to the development of a register based statistical system as follows:

- Legal base
- Public approval
- Unified Identification Systems
- Underlying reliable registers for administrative purposes

Not all countries undertaking Census modernisation satisfy these preconditions to Nordic style Census taking. As such, new approaches to circumnavigating or negating these preconditions are emerging as part of Census modernisation programs.

Legal base

The necessary legislation must be in place to enable administrative data sources to be easily used for statistical purposes. Even if the necessary legislation is not in place, it may still be possible to undertake Census modernisation. However, any constraints in the legislation will carry risks. For example, where data protection legislation constrains linkage or data integration, innovative solutions can be developed and implemented to help address these constraints. The ONS implemented a system of data integration using encrypted match keys, while the Census in Germany in 2011 required the creation of a temporary register due to legal restrictions around data linkage. Countries may also seek to implement new legislation where it does not exist. Examples of new legislation to support Census modernisation comes from the UK Digital Economy Act 2017 ONS UK (2017) and the Swiss Census in 2011 (Schwyn and Kauthen, 2009).

Public approval

It must be acceptable to the general public that the statistical office can use administrative data sources for statistical purposes. The statistical office must have the trust of the public. This not only relates to trust with administrative data but also trust in developing, implementing and explaining new Census methods based on administrative data. There is also a counter argument here. Many countries have and are finding it increasingly difficult to carry out a traditional Census where every person is directly enumerated. Examples of such difficulties in the past include the Netherlands and Germany. Increasing difficulties in getting people to respond to Censuses in the Netherlands led to a Virtual Census in 2000 (Nordholt, 2005). In West Germany, debates about privacy rights led to a boycott of the 1983 Census and its postponement until 1987 (Scholz and Kreyenfeld, 2016).

Unified Identification Systems

Having official identifiers for persons, businesses and property facilitates easy linking of data. It is possible to create linking keys in the absence of such identifiers but it is more laborious, time consuming and prone to linkage error. ONS (ONS UK, 2013) and SNZ (O'Sullivan, 2015) have developed linking methodologies to link administrative data in the absence of identifier keys for persons and properties. Germany compiled a list of all dwellings that existed on Census day called the AGR to enable data linkage between each of the different data sources (Bechtold, 2016).

Underlying reliable registers for administrative purposes

The primary purpose of the registers arises out of the functioning of a society and the development of the underlying administration (social security, taxation, education, health, etc.). Registers and datasets compiled from administrative data sources can suffer from coverage issues. For many countries that have long established register based systems, the official or registered population is the population. However, in considering a more harmonised definition of the population, these countries now accept that coverage errors may arise, even if negligible in size. Netherlands (Gerritse et al., 2016), Switzerland (FSO, 2015), Sweden (Bengtsson and Rönning, 2016) and Spain (Argüeso and Vega, 2014) are examples of countries that have evaluated coverage errors in the registers and found them to be negligible. Israel has used a significant coverage survey to adjust for coverage errors (20% of households) (Kamen, 2005) and anticipates the continued use of a coverage survey to correct for significant coverage errors in the population register. In building population spines or SPDs from administrative data sources, both SNZ (Statistics New Zealand, 2016) and ONS (ONS UK, 2017) anticipate the use of surveys to correct for coverage errors.

Some of the essential features of a Census may also take on a different form in a new modernised Census.

Individual enumeration

Having every person enumerated in their dwelling of usual residence is important to enabling statistics to be easily produced in a coherent manner. With the traditional Census models the concept of individual enumeration is increasingly being challenged. Increased non-response and the requirement to rely more on coverage adjustment factors brings added complications to the traditional Census model. In new Census models, full individual enumeration may not always be possible or easy to produce. Statistics Netherlands gathers Census attributes through existing surveys and deploys complicated methodologies in trying to produce the most coherent set of cross classified statistics possible for the official population (Nordholt et al., 2014). In the 2011 Spanish Census, where a 10% sample was used to collect the attributes, INE accepts that there will be inconsistencies between the Census counts generated from the WCF and the cross classified attribute information generated for the population using the survey (Argüeso and Vega, 2014).

Simultaneity

In the traditional sense, everybody is counted at the same time (i.e., Census night) in order to prevent overcount through double counting. In practice, this translates to a short period of time but with a reference to a particular night. This feature also has a strong presence in different population definitions - i.e., defacto, dejure, usually resident and present. The increasing demand for more regular Census like population estimates at a small area level is going to require increasing use of administrative or secondary data sources. SoL in administrative data sources is emerging as a mechanism for adjusting for coverage issues and producing population estimates in a number of countries - United Kingdom, New Zealand, Israel, Italy (Gallo et al., 2016). SoL are typically observed for an individual over a period of time, possibly a calendar year, and as such there may be a demand among NSIs to broaden the reference period in the population definition as more and more countries modernise their Census. Lanzieri (2013) considers these issues and proposes a population concept based on the *annual resident population* to better facilitate international comparability of results. This concept is based on the amount of time a person resides in a country in a particular calendar year. The French rolling Census, described in Durr (2005), does not directly contain the simultaneity feature but a moving sample that covers the population over a 5 year period is taken as meeting the requirement (UNECE, 2014).

Universality

Universality requires the counting or benchmarking of the population to include every person residing or present in the defined territory of the country at a defined point or period in time. The enumeration provided by the Census should also be validated with an independent coverage check. This feature relates to geography and is a key foundation stone of the Census in both the traditional and modernisation sense. Methods based solely on administrative data risk the exclusion of small groups of the *unofficial* population, a particular group of attention are migrants and unofficial workers as identified in Israel (Blum and Feinstein, 2017), Netherlands (Gerritse et al., 2016; Statistics Netherlands, 2016) and Spain (Argüeso and Vega, 2014).

Small area

A Census is required to be able to produce statistics on the number and characteristics of housing and persons for small areas within the country. A Census must have the capacity to build a high quality address list or register with associated geo-spatial information that will allow each person being enumerated to also be geo-referenced. For traditional Censuses, these address frames can be validated and improved by field staff involved in the enumeration. Coverage surveys that depend on high quality address frames are a key component of validating and calibrating population estimates for those countries that have recently moved to using administrative data. Spain (Argüeso and Vega, 2014), Israel (Kamen, 2005), Germany (Bechtold, 2016) and Switzerland (FSO, 2015) are such countries. The coverage surveys that UK (ONS UK, 2017) and New Zealand (Statistics New Zealand, 2016) are anticipating will also require high quality address frames. These coverage surveys are also critical to addressing *universality* in the modern Census.

Defined periodicity

The Census is also required to be taken at regular intervals to ensure comparable information is available over time. A significant criticism of the traditional Census is that it is typically conducted every 10 years with only some countries conducting one every 5 years. Census modernisation, if done efficiently, offers the possibility of more frequent Census type statistics at significant geographic detail. The European Statistical System (ESS) is considering the possibility of producing annual Census like statistics for reference year 2024 onwards which can only be achieved with acceptable costs through the use of administrative data.

A UNECE task force, working on recommendations for the use of registers and administrative data for population and housing Censuses (Nordholt, 2017), proposed a common framework for *register based* and *combined* Censuses with 5 key stages covering

- data sources,
- linkage and transformation,
- creation of statistical registers and population datasets,
- quality measurement,
- assurance and outputs for dissemination.

Census modernisation can be summarised as, first, creating the necessary statistical population and housing datasets from available data sources, and then validating or correcting for errors in the datasets with respect to the target population. The different types of errors in the statistical population datasets can typically be described as linkage errors, domain misclassification errors or coverage errors. Linkage errors will occur where there is an absence of official identifiers to enable high quality deterministic linkage, e.g. Germany, UK and New Zealand. Domain misclassification errors will occur due to incorrect or conflicting attribute information (e.g. date of birth, gender, address) being recorded in the underlying data sources. Coverage errors occur due to a mismatch between the statistical datasets and the target populations and can include undercoverage, overcoverage or both. For some countries, including the Nordic countries, Netherlands, Switzerland and Spain, it is generally accepted that the official population registers are sufficiently aligned with the target population that it isn't necessary to implement additional methodologies to correct for errors - a simple count of records on the registers is sufficient to create Census like population estimates. When this is the case, there is a significant added advantage in that the underlying methodology is easily explained to users. Where significant correction of error is required, it appears that a rules based approach, such as investigated by Sweden (Bengtsson and Rönning, 2016), will not be sufficient without validation of those rules through some other mechanism. This gives rise to the requirement for a coverage survey as being proposed by ONS and SNZ in their Census transformation plans to correct or validate estimates.

The coverage surveys deployed appear to be modelled on the typical Census Coverage survey where linking of records is undertaken by person within household in a sample of small areas between the two lists, SPD and coverage list. Estimates of overcoverage and undercoverage are compiled for different population groups in the small areas and these estimates are then applied to similar small areas and groups in the form of weighting or imputation to adjust or correct population counts from the SPD. Examples of coverage surveys are found in Switzerland (FSO, 2015), Germany (Bechtold, 2016), Israel (Kamen, 2005) and Spain (Argüeso and Vega, 2014).

1.2.2 Census of Population in Ireland

1.2.2.1 One evening in 1991 ...

One early Thursday evening in December 1991, the then recently retired Director of the Central Statistics Office (CSO), Thomas P Linehan, read his paper (Linehan, 1992) on the *History and Development of Irish Population Censuses* before the Statistical and Social Inquiry Society of Ireland. This was a very topical paper as 1991 was a Census year. While the paper makes for an interesting and factual read, the ensuing discussion was set up very neatly when the paper concluded with the lines:

> "Perhaps it is time to look at the long-term plans for all aspects of future Censuses, not only the content but the collection and processing as well."

The audience included the Census management team at the CSO along with key users and demographers. The following quotes from that discussion provide insights into thinking with respect to the modernisation of Censuses in Ireland at that time.

> "I cannot at this stage see how individuals in households throughout the country would complete their Census forms electronically and send the results *over the wires* rather than complete the paper copy. It would be a brave person who would introduce such a change given the success and stability of the present methods and given the difficulties of ensuring quality control in such an environment."

"... the so-called Census undercount question is a major item of concern in the US ... Thankfully so far in Ireland, we have not had to wrestle with a Census undercount problem. A lot of credit must go to the very professional field forces ..... Their diligence and persuasiveness in following up to the small number of difficult cases has contributed in no small measure to the success of the Irish operation."

"The use of population registers for determining population counts is not a practice to which we are accustomed in Ireland. .... the nearest we have to such a system in Ireland was the population register introduced during the *Emergency* where on the basis of enumeration each individual was accorded a unique number for the issue of a ration book. However the unique reference number disappeared with the *Emergency* and in my view, given the Irish psyche we will probably have to wait for another Emergency before we have the appearance of unique reference numbers."

"... concerns the processing of Census results. While traditional paper questionnaires persist the problem of transferring their contents into machine readable form will be with us. However, this is not a major problem. Regardless of what medium we use much resources will have to be invested in ensuring the integrity of the data base .....This scrutiny could be achieved more quickly and earlier publication of the main results put in place if a greater block of resources were allocated for a correspondingly shorter period of time. While we were not successful in the recent past in having this view accepted by our colleagues in the Department of Finance ..... from a purely cost benefit point of view it does not make good economic sense to invest over £6m in the fieldwork phase of the Census and then to skimp on resources afterwards so that the final results are delayed."

"It strikes me, the primary function of a Census notwithstanding, that there was more concern with the type of individual remaining in the country rather than the numbers that left. This is indicated by the fact that emigration data was collected in just one Census while information was systematically collected on lunatics and persons in prisons in all Censuses until 1911. This observation probably stems from a conspiracy theorist's point of view but unfortunately it may have equal validity in the present."

"I have previously expressed concern about this country's lack of information about the number of and profile of its emigrants. This is especially disturbing as migration rather than natural increase in population has consistently been the primary determinant of demographic change in Ireland.

> Unfortunately, the Census of Population has been unable to provide a solution to this problem ... The ideal solution to this kind of problem would be recourse to a form of central population register such as exists in some of the Scandinavian and Benelux countries."
>
> "The Census of Population is mostly derived from their population register. Such a system would, unfortunately, prove extremely difficult to effect in Ireland for historical, cultural and possible even constitutional reasons."
>
> "Given the value of such *usual resident* data and the prevalence of *de jure* Censuses of Population in many other countries which reflect the usual or normal family and / or household situation I conclude by asking whether the *de facto* Census of Population, which has existed in Ireland since 1841, should be retained in preference to the *de jure* format?"

So in 1991, the Census of Population was the key vehicle for providing population estimates in Ireland. It was conducted in the traditional way providing a *de facto* count of the population on a given night. The Census of Population provides a hugely rich source of all types of information on the Irish population. It was a big logistical operation employing a field staff of 3,200 enumerators at a cost of over £6m (Irish pounds). It took two years to process the data. The Census of Population is a tried and tested operation repeated at regular intervals (usually every 5 years). To introduce change to the way the Census is undertaken would involve introducing significant risk to a tried and tested procedure. However, the Census of Population did not supply all the answers. It could only provide limited information on immigrants with respect to the year it was conducted and could not provide information on the sizeable number of persons that were emigrating abroad at that time. It used a *de facto* formula as opposed to a *de jure* formula used in many other countries. In 1991, the Census relied on the integrity, tenacity and ability of its field staff to eliminate any undercount problem.

Ireland had previously undertaken such a register based population count in 1941 during World War 2 where unique identification numbers were used to identify ration book holders. These numbers disappeared with the discontinuation of the need for ration books. In 1991, the introduction of a CPR was considered fanciful as it would be difficult to effect due to cultural and legal considerations. It was thought that Ireland could not emulate Scandinavian or Benelux countries in this regard.

1.2.2.2 25 years later ...

The Census in 2016 was a traditional Census with a field force of 5,000 enumerators dropping off and picking up 1.9 million Census questionnaires to households in the State.

Census 2016 preliminary results were published in April 2017 (CSO, 2017) showing that approximately 4.76m persons were present in the State on Census night (24th April 2016). The Census cost approximately €60m. No internet option was offered for the return of forms. There is a continued reliance on the field force and the underlying collection systems to eliminate any undercount problem.

The primary difference in the processing of Census 2016 and Census 1991 is that the manual keying requirement for Census forms has been removed and replaced by a system that can scan forms and use Optical Character Recognition (OCR) technologies and read the value of each field in the form. Other significant changes include the use of internet technology for dissemination and an enhanced use of geospatial information to enable more geographical detail.

In terms of Census modernisation, it would appear that little has changed since 1991. However, look beyond the Census and, in particular, at how the Irish Statistical System (ISS) is developing and an opportunity to transform Census taking in Ireland can be identified. The author has previously documented this unfolding opportunity (Dunne, 2015).

1.2.3 Irish Statistical System (ISS)

The groundwork for the development of the ISS was laid with the coming into effect of the Statistics Act (1993) (Statistics Act, 1993). The Statistics Act envisaged the ISS as a register based statistical system. The further development of the ISS was delayed as the CSO moved from Dublin to Cork as part of a radical decentralisation program. This move had a significant impact on the development of the CSO and therefore on the development of the Irish Statistical System.

It wasn't until the early 2000s that the CSO started engaging with other Public Sector bodies and produced a number of reports (CSO, 2003, 2006, 2009) that would help kick start the development of the Irish Statistical System. Furthermore, in 2009 CSO consolidated all its activities relating to administrative data for statistical purposes and the development of the Irish Statistical System into the new Administrative Data Centre (ADC). The operation of ADC and its environment is described in more detail by Hayes and Dunne (2012).

Building on the work done with the Irish Revenue Commissioners (CSO, 2009), CSO for the first time profiled its business register solely from administrative data sources and produced a first comprehensive business demography product for reference year 2008 in 2010.

In late 2011, the National Statistics Board (NSB) produced a position paper (NSB, 2011) on the value of joined up data for joined up Government that managed to put data at the heart of Public Service Reform (DPER, 2011). For the first time, the development of a National Data Infrastructure (NDI), based on the collection of Official Identification Numbers for persons, businesses and property for any transactions with the State, was advocated. The NDI takes inspiration from the ideas originally developed by Nordbotten (Nordbotten, 2010) in the 60's and developed into a simple model for a socio-demographic statistical system by Thygesen (Thygesen, 2010) in the 80's. These ideas underpin the register based statistical systems in operation in Scandinavian countries today. The rationale for an NDI in Ireland is presented by Macfeely and Dunne (2014).

In the meantime, the use of a Personal Public Service Number (PPSN) has become increasingly more common. The PPSN in Ireland is the official Person Identification Number (PIN) and is the responsibility of a special unit in the Department of Social Protection (DSP). This number originated as the Revenue Social Insurance (RSI) number in 1979, when Social Welfare Services and Revenue integrated their person identification number systems. In 1998 it was renamed the PPSN and today is used across many public administration systems. Anybody living in Ireland and entitled to engage in a transaction with the state (tax, education, welfare, health etc.) is generally required and entitled to obtain a PPSN. A PPSN is typically assigned to a person when their birth is registered and enables a mother to receive an ongoing universal child benefit payment in respect of their child. More information on the PPSN is available from the DSP website (http://www.welfare.ie/en/Pages/home.aspx).

The CSO, in advocating for the development of the ISS and the underlying NDI, collaborated on or developed a number of projects that utilised the PPSN in demonstrating the power of joined up data. These projects typically leveraged the longitudinal possibilities and the capacity to link across multiple data sources using a PIN. Examples of such projects included exploring the dynamics in the labour market (Dunne, 2011) and where do school leavers go (DES, 2013a,b). The first example allowed the CSO to react quickly to a huge demand for information on jobs after the economy plummeted in 2009 while the second example provided comprehensive information on school leavers and eliminated the need for a costly and complex survey. These type of projects served to promote the need for an NDI in Ireland.

The primary area for development in the NDI is the use of official identifiers for properties. Ireland is probably unique among developed countries in that it did not have a postcode system up until 2015. In 2015, Ireland introduced the EIRCODE system with a postcode for each letterbox in the country. This postcode system will prove invaluable with respect to geospatial referencing capabilities of the NDI and in particular the ability to link across multiple data sources based on address. In the absence of this postcode system, the NDI faces the challenge of trying to integrate data based on address strings where address strings are not standardised. Furthermore, in 35% of cases the address strings are not unique and require the person's name to also be attached to the address to ensure post is delivered. The EIRCODE system utilises the database that postmen in

Ireland use to deliver the mail. To date, there has been significant take up by the private sector including internet mapping applications such as Google Maps. It is now a priority for the CSO and the NSB to promote and agitate for the uptake of the EIRCODE on the databases underpinning public administration systems.

The message from the NSB is that the use of official identifiers is primarily for effective and efficient public administration. Statistics and living in a more informed society is a downstream benefit. The NSB has outlined and published its vision in its most recent Statement of Strategy (NSB, 2015).

1.2.4 The Emerging Census Opportunity and Consideration of Preconditions for a Register Based Census

The Statistics Act, 1993 provides the necessary legal base.

The CSO has, for some time, been using administrative data sources in the compilation of Official Statistics and has a strong track record in this regard. There exist strong identification systems for businesses and persons in the state.

The business identification system is primarily governed by the Revenue Commissioners and shared with the CSO. The PPSN or PIN system has more widespread adoption across Government bodies and the master list is maintained by the Department of Social Protection (DSP). The address identification system, EIRCODE, is still in its infancy but is gathering momentum in its uptake across Government.

There are strong underlying public administration systems. While no CPR exists it is possible for CSO to create a Statistical Population Dataset (SPD) using the underlying administrative data systems as satellite registers. The SPD, with ideally one record per person, takes the same role as a population frame created from the CPR.

As the EIRCODE develops and becomes more commonplace, the ability to create an enhanced SPD with high quality linkages between persons and dwellings becomes more feasible.

This is an emerging Census opportunity and is closely linked with the ability to link persons to dwellings.

One of the first milestones in modernising the Census is developing the capability to compile reliable population estimates at State level from administrative sources. This milestone does not necessarily have to rely on linking persons to dwellings. It does however rely on the ability to identify those resident in the State for a given reference point or period.

1.3 PECADO - The Simple Idea

1.3.1 If You Don't Have a CPR, Build an SPD

Many countries do not have a CPR. Some of these countries are now actively considering how to get the benefits of a statistical system based on registers. In the absence of a CPR, the simple idea is to compile a statistical register or SPD using available data sources.

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 must be living in the State. In compiling an SPD from multiple data sources, 4 main types of error need to be dealt with in order to cover a target population:

- Overcoverage: Where the SPD has units that do not belong to the target population.
- Undercoverage: Where the SPD is missing units that belong to the target population.
- Linkage error: Where units are incorrectly identified as other units, for example where a PIN is incorrect.
- Domain misclassification: Where an attribute has an incorrect value for a unit. This may occur when the same or similar attributes on different contributing data sources have conflicting values.

Overcoverage

First attempts to create an SPD from multiple sources typically focus on registration information, i.e., persons registered for health, tax, social welfare etc. For many persons residing in the State there is an incentive to register with various public administration systems (to obtain the benefits) while there is little or no incentive to de-register. Hence, the approach of focussing on registrations will typically lead to overcoverage errors. It is very difficult to eliminate over coverage through reliance on registration information alone unless the registration systems also include a strong incentive for persons to also de-register. Overcoverage problems are typically addressed using an overcoverage survey.

Undercoverage

Another challenge is that it may not be possible to find enough administrative data sources to cover the target population when creating an SPD. Although, when considering all aspects of life, from the cradle to the grave, it is difficult to identify sizeable groups that would not interact with Public Services in some way.

Linkage Error

Another challenge for those countries that don't have a CPR and don't have a high quality PIN to link persons, is one of linkage error. In such situations, linking becomes dependent on being able to create high quality matching keys using demographic information such as date of birth, name, place of birth and address. ONS, SNZ and Statistics Canada face these challenges and have done a lot of interesting work with probabilistic and deterministic matching using derived match keys in the absence of a PIN. There is a greater risk of linkage error with probabilistic matching and deterministic matching with derived match keys than there is with the use of a PIN and deterministic matching. If a PIN is universally used with complete accuracy then linkage error is eliminated. There is also a risk of linkage error where the PIN is not universally used across all relevant data sources.

Domain misclassification

Another type of error is domain misclassification or attribute error. This type of error occurs when an incorrect attribute value is recorded against the statistical unit, for example a person identified as being male when, in fact, the person is female. While, generally, we accept such misclassification as random errors for statistical purposes, it is important to be aware of any incentives that may cause such errors to contain a systematic bias. One interesting area to note is the rules with respect to age in the underlying programmes that generate the administrative data. For example, in considering those looking to receive a State Pension there maybe an incentive to lie and say one is older than one is to receive pension entitlements before time - an entertaining example of this is described by Ó Gráda (2000) and tells the story about when the State Pension was first introduced in Ireland in 1909. The birth records did not go back far enough and, in order *to counteract the tall tales being told by relatively young and sprightly persons*, a system was set up whereby a persons age could be verified against the 1841 and 1851 Census records. However, this may also have had an adverse effect on how a person reported their age in the 1911 Census. Another regular source of domain misclassification is where there is conflict between two data sources that have common attributes with values that don't agree. A decision needs to be made on which data source to choose for the correct attribute and this can be done at a unit level by prioritising data sources or on some other basis (eg some probability basis).

The CSO, like other NSIs, 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 PPSN 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 availability of a high quality PPSN on administrative data sources enables users to use deterministic matching with a high degree of confidence. The master file of PPSNs 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 when linking or modelling data on these attributes.

The *PAR* has taken a different approach to initial attempts at building SPDs in 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 problem types 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 (SoL)* approach and only uses the registration information from the master file of PPSNs to provide consistent attribute information such as date of birth, nationality and gender. This SoL based approach can be summarised as only including persons 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 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 with respect to domain misclassification are eliminated or minimised by using the original master list of PPSNs as the truth for age, gender and nationality attributes. 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 remaining problem of any consequence to be dealt with is one of *undercoverage*.

Following is an overview of the administrative data sources (transactions) included in the *PAR*.

Childrens Benefit: Universal payment made on behalf of each child, generally to the mother, while the child is under 18 and in full time education. Indicators are used for both the mother and the child.

Early Childhood Care: Each child is entitled to 1 year paid childcare prior to attending primary school.

Primary Online Database (POD): Student enrolments in primary education in the State. Typically for children aged 5 to 12 years.

Post Primary Pupils Database: Student enrolments in secondary education. Typically for children aged 12 to 18 years.

Higher Education Enrolments Database: Student enrolments in third level education. Typically for children aged over 18 years.

Further Education Awards Database: Student awards in further education (excluding higher education). Typically for persons aged over 16 years.

Employer Employee Tax Returns: A database of paid employees (including occupational pensions) created from the employer returns to the Irish tax authorities each year.

Income Tax Returns: Tax returns filed by persons for any taxable income other than paid employments each year.

Social Welfare: Social welfare payments to recipients each year.

Primary Care Reimbursement System (PCRS): Part of the public health system in Ireland where those who qualify are entitled to contribution or payments towards health care. A number of schemes are included and qualification typically depends on a number of factors - age, health condition, income, to name a few. In 2016, over 2 million people qualified for some type of benefit or refund.

State Pension: All those entitled to a State Pension on reaching retirement age.

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 SoL 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 SoL activity is admitted as evidence from the corresponding source only if the PIN can be identified.

Building the PAR in this manner implies that it may and probably does contain undercoverage with respect to the target population, however problems with overcoverage, domain misclassification and linkage error are eliminated or minimised to such an extent that they are negligible in comparison to undercoverage.

Direct counts from the *PAR* will therefore need to be adjusted for undercoverage errors if they are to be used as population estimates.

1.3.2 Now Adjust SPD counts for Undercoverage to Obtain Population Estimates

1.3.2.1 Methodology background

In order to compile population estimates from the PAR, direct counts will need to be adjusted for undercoverage.

In adjusting the PAR for undercoverage we look to the traditional approach for adjusting for Census undercount. Adjusting for Census undercount typically involves undertaking an undercoverage survey (UCS) to generate recapture data and adjusting using capture recapture methods. These methods are introduced and covered in a number of Statistical text books (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 Dual System Estimation (DSE) methodologies in Official Statistics. The Petersen Model, described by Wolter (1986), is the starting point in chapter 2 for our consideration of a model to adjust for undercoverage. We explore this model and end up with a DSE model with more relaxed assumptions. In particular, the DSE model only requires each unit in the population to have equal probability of being caught in one list. This assumption is often referred to as the *homogeneous capture assumption*. We denote this list satisfying this assumption as list B and the list that is not required to satisfy the assumption as list A.

1.3.2.2 List B - Adjusting for Undercoverage

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 categories such as learner driver licences and bus or truck licences will need to be renewed more often, 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. It is assumed that those that do not hold a driving licence will behave in the same way on public administration systems as those that do hold a driving licence.

The list of those persons that renewed their driving licence or applied for a new one in the relevant calendar year is proposed as a suitable list B candidate for adjusting for undercoverage on the SPD. This list B will be denoted as the driving licence dataset (DLD). Historically, a person did not require his or her PIN to obtain or renew a driving licence. However, since 2013 the provision of a verified PIN has become mandatory. Again, a person is included in the DLD provided only the PIN 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. For practical purposes, this population concept is equated to the usual resident definition used as part of the population concept in the Official Census. For statistical purposes, a person is considered usually resident if they reside or intend to reside in a country for a period of 12 months or more[1].

It is assumed that those that do not hold a driving licence will behave in the sameway on public administration systems as those that do hold a drivers licence.

To better understand the closed population assumption, where list A and list B are both drawn from the same population, we will describe a number of population concepts.

Census Night Population (U_I): This is the *de facto* definition currently of the Irish Population Census. It includes every person that is in the State on a given date, regardless of the status or nature of the presence.

Usually Resident Population (U_{II}): While the exact definition of usually resident status may differ, the concept is typical and in principle feasible for register-based population counts. In the countries that have implemented completely register-based Census, the usually resident population also has a specific reference date.

Hypothetical SPD Population (U_A): This comprises any person who has had or *in principle* could have had interactions with the relevant public administrative systems *during* a calendar year. The inclusion of the latter is necessary because the PAR is not a population register.

Hypothetical *DLD* Population (U_B) This comprises any person who holds or *in principle* could hold an Irish driving licence. The latter is necessary in order to make the DLD relevant for population size estimation. Otherwise the actual DLD population could be enumerated directly.

[1] 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

Under the closed population assumption,

$$U_A \subseteq U_B = U_{II} \tag{1.4}$$

that is the hypothetical SPD population is a subset of the hypothetical DLD population which equates to the usually resident population in the state. We note that, in practice, U_B doesn't include those of an age that do not qualify to apply for a driving licence, therefore, the assumption is only valid across those age groups that qualify to hold a driving licence..

Blocking both list A and list B by single year of age, nationality grouping and gender will also relax the *homogeneous capture* assumption so that it only has to apply within blocks. This allows for likely different propensities to hold a driving licence across age, gender and nationality groupings. In other words, age, nationality and gender are treated as covariates as part of a strategy to deal with heterogeneous capture rates across age, gender and nationality.

Blocking in this manner also facilitates easy disaggregation of population estimates by these same groupings.

1.3.3 The Idea in Practice

In theory, it now looks feasible to compile a system of population estimates for Ireland using only administrative data sources. But if we are going to use this system of population estimates for Official Statistics the robustness of the system will need to be demonstrated. We refer to this system as the PECADO system of population estimates.

Chapter 2 develops and presents the DSE methodology to be used in compiling population estimates from administrative data only. The chapter also develops and presents Trimmed Dual System Estimation (TDSE) as a tool for the hunting of overcoverage or erroneous records and for the evaluation of underlying data sources contributing to the SPD. The chapter also explores what happens when the *homogeneous capture assumption* doesn't hold up.

Chapter 3 explores the PECADO system of population estimates to see how robust it is. In particular, we use the Trimmed Dual System Estimation (TDSE) toolkit to look for erroneous records in list A and consider an alternative source as list B to investigate if the assumption of *homogeneous capture* holds for the dataset of driver licence renewals and applications. We also consider the Census counts from 2011 and 2016 and compare them with the PECADO population estimates.

Chapter 4 considers an administrative data list in a DSE setup to estimate Census undercoverage. In particular, we apply the methods in the 2016 Census setting.

Chapter 5 considers how the proposed PECADO system can be extended to also estimate for gross population flows. A good system of population estimates will by definition provide for a good system of net population change estimates (simply, the difference between two sets of population estimates). However, there is also still a requirement to estimate gross population flows; that is, gross inflows (births plus immigration) and outflows (deaths plus emigration).

Chapter 6 summarises our conclusions and proposes the adoption of a new system of Population Estimates Compiled from Administrative Data Only (PECADO) for Ireland.

Chapter 2

Methodology - The PECADO toolkit

2.1 Introduction

In this chapter we explore DSE methods and adapt them for use in the PECADO project. We label this refined toolkit, the PECADO toolkit to differentiate between the DSE methods as applied in the project and those that are more traditionally applied in Census Coverage Surveys (Wolter, 1986).

Section 2.2 reviews the traditional Petersen Model as presented by Wolter (1986) before presenting an alternative derivation of DSE methods that allows for more relaxed assumptions. This alternative derivation of DSE methods is used in the PECADO project/toolkit, as the relaxed assumptions facilitate the application of the methods in a much broader context.

In section 2.3, we present an extension of the DSE methodology, which we call Trimmed Dual System Estimation (TDSE), that allows for the hunting of erroneous records in list A in the DSE setup. This is an important methodological innovation as it now allows DSE methods to be applied in situations where overcoverage exists, provided an effective strategy to evaluate list A and trim those parts of list A that contain erroneous records using TDSE can be deployed.

In section 2.4 we explore the impact of violating the assumption that each population unit has an equal probability of being caught in list B and discover that this assumption can also be further relaxed in certain situations. We also present a test for determining whether a violation of this assumption results in a significant bias by comparing a stratified estimator with a simple estimator (without stratification).

		list B		
		in	out	
list A	in	p_{i11}	p_{i12}	p_{i1+}
	out	p_{i21}	p_{i22}	p_{i2+}
		p_{i+1}	p_{i+2}	1

Table 2.1: Multinomial distribution ϕ_i as per the Wolter's *Multinomial Assumption* (Wolter, 1986)

2.2 Dual System Estimation methodology

2.2.1 Dual System Estimation (DSE) Methodology - Traditional Petersen Model

Wolter (1986), summarises a number of variations of capture recapture models for estimating undercoverage in Census data. We use Wolter's setup to consider the Petersen model.

Consider a population U of unknown size N. A Census is conducted to enumerate every i^{th} person in U. For various reasons the Census will fail to enumerate every person and provides an undercount of N. To produce an estimate of the undercount and therefore the population size an additional sample survey of the population is undertaken. The list of persons enumerated in the Census is often referred to as list A, while the list of persons included in the sample response is referred to as list B.

The Petersen model, also known as the Dual System Estimator or Lincoln Index, then requires the following assumptions.

1. **The Closure Assumption** The population U is closed and of fixed size N.

2. **The Multinomial assumption** The joint event that the i^{th} person is in list A or not and in list B or not is modeled by the multinomial distribution ϕ_i with parameters as outlined in table 2.1.

3. **Autonomous Independence** List A and list B are created as a result of N mutually independent trials using distributions ϕ_1, ϕ_2, ..., ϕ_N. The cell counts, x_{11}, x_{12} and x_{21} from the N trials in table 2.2 are considered observable. Cell count x_{22} and population size N are unknown and need to be estimated based on the model.

		list B		
		in	out	
list A	in	x_{11}	x_{12}	x_{1+}
	out	x_{21}	x_{22}	x_{2+}
		x_{+1}	x_{+2}	$x_{++} = N$

Table 2.2: Cell counts from N mutually independent trials under Wolter's *Autonomous Independence Assumption* (Wolter, 1986)

4. **The Matching Assumption:** There are no errors, through either omission or inclusion, in determining the match between list A and list B.

5. **Spurious Events Assumption:** All erroneous records due to spurious events are removed from both list A and list B prior to estimation.

6. **The Nonresponse Assumption** In the context of a coverage survey being used as list B, it is assumed that there is enough information about non-respondents to permit exact matching between the survey (list B) and the Census (list A). Surveys and Censuses typically contain non-response.

7. **The Poststratification Assumption** Any variable used for post stratification (i.e. age and sex) is correctly recorded for all persons on both list A and list B.

8. **Causal Independence** The event of being included in list A is independent of the event of being included in list B such that the cross product ratio θ_i satisfies equation 2.1.

$$\theta_i = \frac{p_{i11}p_{i22}}{p_{i12}p_{i21}} = 1, \text{ for } i = 1, ..., N \tag{2.1}$$

9. **Homogeneous Capture within List Assumption** The capture probabilities satisfy $p_{i1+} = p_{1+}$ and $p_{i+1} = p_{+1}$ for $i = 1, ..., N$ where p_{1+} is the probability of list B persons being captured in list A and p_{+1} is the probability of list A persons being captured in list B. (This assumption is numbered 11 in Wolter's paper.)

Under these assumptions and using Maximum Likelihood methods, Wolter shows that an estimator for N is given by equation 2.2 with a variance estimator given by equation 2.3

$$\hat{N} = \frac{x_{1+}x_{+1}}{x_{11}} \tag{2.2}$$

$$\hat{V}\left[\hat{N}\right] = \frac{x_{1+}x_{+1}x_{12}x_{21}}{x_{11}^3} \tag{2.3}$$

In considering Wolter's assumptions, we summarise them and restate them into the following 6 assumptions

- The Closure Assumption: As per Wolter.
- The Matching Assumption: As per Wolter.
- No erroneous records: There are no erroneous records in either list A or list B. This combines a number of Wolter's assumptions (*5. Spurious Events Assumption*, *6. The Nonresponse Assumption.* and *7. The Poststratification Assumption.*)
- Causal Independence Assumption: As per Wolter.
- Homogeneous Capture within List Assumption: As previously per Wolter.
- Independent Capture within List Assumption: The event that a person is captured in a specific list (A or B) is independent of the event of any other person being captured in that list.

The *Independent Capture within List Assumption* when considered in conjunction with the other restated assumptions adequately captures Wolter's *Multinomial* and *Autonomous Independence Assumptions.*

2.2.2 Dual System Estimation Methodology - Adjusted

We take as our starting point Zhang and Dunne (2018), and following this approach we have:

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 $B \subset U$.

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 is *The Matching Assumption.*

The notation used has been changed in places from that of Wolter due to the differing assumptions that are made with respect to lists A and B. x_{1+} has now been replaced by x, x_{+1} has now been replaced by n and x_{11} has now been replaced by m.

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 $i \in U$. We shall refer to this as the assumption of *homogeneous capture* (of list B). This equates to the *Homogeneous Capture within List Assumption* stated earlier but only for list B. It is the starting point of the development of the estimator. Heterogeneous capture can be accommodated through post stratification to ensure that the homogeneous capture assumption holds within each 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 moments (MoM) estimator, given by

$$\hat{N} = nx/m \tag{2.4}$$

We call this a MoM estimator as A is treated as fixed with size x, but m can vary in size depending on the outcome of which n persons have been chosen from the population, and as such, the estimator is based on the list B sample of size n chosen.

Developing the DSE in this manner requires the following 3 assumptions:

No erroneous records: A closed population ensures no records from outside the population but we also suppose there are no duplicate records or incorrectly identified records in either list A or list B.

Matching assumption: There is no linkage error when matching records between list A and list B.

Homogeneous capture with respect to list B: Every unit i in the population U has an equal chance π of being captured in list B.

These assumptions are more relaxed than those described in Wolter (1986). With respect to Wolter's assumptions we now only need to retain the homogeneous capture assumption with respect to list B. This allows a much broader application of DSE particularly when list A is compiled from administrative data sources where it is generally difficult to justify the argument of homogeneous capture for that list. The development of the DSE here also negates the multinomial assumption arising when cross classifying list A and list B as outlined by Wolter.

The variance of $\hat{N}$ is obtained as follows: List A with x records is treated as fixed and an extra assumption of *Independent Capture* is made such that $V[n] = N\pi(1-\pi)$ and $V[m] = x\pi(1-\pi)$.

Now also let $n = m + n_{A^c}$ where n_{A^c} is the number of population units that are not in list A but are enumerated in list B. Provided there is *Independent Capture*, 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 have

$$\hat{v} = \widehat{V}\left[\hat{N}\right] = \frac{n(n-m)x(x-m)}{m^3} \tag{2.5}$$

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 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 typically 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. 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.

The methodology presented in this section underpins the compilation of population estimates in Chapter 3, the methodology to explore undercoverage in the traditional Census using an administrative data source in chapter 4 and forms a core part of the proposed methods to estimate gross population flows in chapter 5.

2.3 Trimmed Dual System Estimation (TDSE) - dealing with erroneous records in list A

2.3.1 Ideal DSE, given erroneous enumeration

We take as our starting point the DSE methods as described in section 2.2.2. Now we relax the assumption that there are no erroneous records. We allow for the possibility that list A has r unknown erroneous records and again develop the methodology in the same way.

Let N be the unknown size of the target population, denoted by U. Let A be the *first* list enumeration that is of size x. Suppose list A is subject to over-counting, and the number of erroneous records is r, i.e. the size of set $\{i; i \in A \text{ and } i \notin U\}$. Suppose list A is subject to under-counting as well, so that $x - r < N$. Let B be the *second* list enumeration 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 we refer to simply as the *matching assumption*. Suppose that error-free matching between A and B gives rise to the matched list AB with m records.

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 $i \in U$. Again, we shall refer to this as the *homogeneous capture assumption* (of list B).

Given the assumption of homogeneous capture, we have

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

Moreover, let $\delta_{iA} = 1$ if $i \in A \cap U$, 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 $\sum_{i\in U} \delta_{iA} = x - r$.

Given the assumptions of homogeneous capture and matching, we have

$$E[m|\boldsymbol{\delta}_A] = (x - r)\pi$$

which is the expectation of the number of records in list AB on applying the constant capture probability π to the $x - r$ records in list A with $\delta_{iA} = 1$. 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

$$\tilde{N} = \frac{n(x - r)}{m} \tag{2.6}$$

Meanwhile, let the naïve DSE, which ignores the erroneous enumeration 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 $r > 0$.

Moreover, instead of the assumption that neither of the two lists contains erroneous enumeration, we now allow for erroneous enumeration in list A, in order to cope with the fact that the underlying administrative sources may contain over coverage or erroneous records. Consequently, 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 target population units in the list B enumeration, 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.

2.3.2 Trimmed DSE

The estimator (2.6) 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.

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

$$\hat{N}_k = n\frac{x - k}{m - k_1} \tag{2.7}$$

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 undercount, 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 (2.6), *regardless* of how systematic the trimming is in removing records from list A. Potential systematic under-coverage 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 are useful in removing erroneous records from list A.

Result 1 (Primary Result):

- 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.
- 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.
- 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.

Given that mk/x is the expectation of k_1 under *random* trimming of k records from the x records in list A, one can expect the trimmed estimate, equation (2.7), to be lower than the naïve estimate nx/m, provided a relatively smaller number of trimmed records are confirmed to be non-erroneous, i.e., they are found in list B. In other words, trimming can be expected to adjust the untrimmed DSE in the right direction, as long as it is more effective at picking out the erroneous records than simple random sampling.

If trimming does not have any impact on the estimator, $\hat{N}_k = \hat{N}_0$, then either there are no erroneous records in list A or the trimming strategy used is no better than randomly trimming records from list A. We consider a trimming strategy as an approach to trimming that attempts to remove all erroneous records from list A. A trimming strategy will typically involve a number of trimming steps.

If trimming results in $\hat{N}_k > \hat{N}_0$ then we have removed proportionately fewer erroneous records from list A than are contained in list A. In practice, this is highly unlikely as in order to reduce bias in the trimmed estimate the trimming strategy will be focused on trimming parts of list A where erroneous records are more likely to occur rather than where they are less likely to occur. However, we include this consideration for completeness.

Result 2: If $k < r$, then $\tilde{N} < \hat{N}_k$.

It is desirable to avoid *under-trimming*, i.e. $k < r$, so that the trimmed DSE, equation (2.7), can converge with the ideal DSE, equation (2.6), with the implementation of an effective trimming strategy.

Proof: We have $(x-r)/m < (x-k)/(m-k_1)$ if and only if $(k-r)/(x-r) < k_1/m$, which is always the case provided $k < r$ since $k_1/m \geq 0$. $\square$

Result 2 states that the trimmed estimate cannot remove all bias due to erroneous records when $k < r$. Therefore, for all bias to be removed, a necessary (but not sufficient) condition is that $k \geq r$. For instance, if it is suspected that 10% of records in list A are erroneous then trimming list A by less than 10% of records cannot remove all the bias. A trimming strategy can be said to be perfect when $(k, k1) = (r, 0)$ in terms of removing bias through trimming.

Result 3: If all the r erroneous records are among the k trimmed ones, then $\widehat{E}[\hat{N}_k] = \tilde{N}$.

Proof: The capture rate in list B of the $k-r$ trimmed non-erroneous records is π, whose estimate is $m/(x-r)$, so that $\widehat{E}[k_1] = (k-r)m/(x-r)$ and $\widehat{E}[\hat{N}_k] = n(x-k)/(m-\widehat{E}[k_1]) = n(x-r)/m = \tilde{N}$. $\square$

To summarise, as long as one is able to trim the erroneous records in list A more effectively than when randomly trimming records, equation (2.7), can be expected to reduce the bias of the naïve DSE and move it closer to the ideal DSE, equation (2.6). If the trimming succeeds in removing all erroneous records, the expectation of the trimmed DSE 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 $\tilde{x} = x - r$. As explained before, we prefer to treat the corresponding list A with $\tilde{x}$ records as fixed. To obtain the variances of n and m, we make an extra assumption of *independent capture*, such that $V[n] = N\pi(1-\pi)$ and $V[m] = \tilde{x}\pi(1-\pi)$. Moreover, let $n = m + n_{A^c}$ where n_{A^c} is the number of population units that are not in list A but are enumerated in list B. Provided independent capture, we have $Cov[n, m] = Cov[m + n_{A^c}, m] = V[m]$. Again, we obtain

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

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

Replacing N by $\tilde{x}n/m$ and π by $m/\tilde{x}$, we have

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

We turn now to the trimmed DSE $\hat{N}_k = x_k n/m_k$, where $x_k = x - k$ and $m_k = m - k_1$. For variance estimation under the same assumptions as those for $\tilde{v}$ above, one needs the number of remaining erroneous records among the trimmed list A with x_k records, which is not known. As an approximate remedy, we propose to make an additional tacit assumption 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

$$v_k = \widehat{V}\left[\hat{N}_k\right] = \frac{n(n-m_k)x_k(x_k-m_k)}{m_k^3} \tag{2.8}$$

If we consider an ineffective trimming strategy where k_1/k is small and many records need to be trimmed before all erroneous records are removed from list A, we see from the variance, given by equation 2.8, that there is a danger that the trimmed estimate will become unstable due to high variance (x_k reduces much faster than m_k).

2.3.3 Stopping rules

Notwithstanding the theoretical assurance above, some practical stopping rules for the trimming are needed that can give an indication of when to stop. Below, three stopping rules are described, all aimed at the same stopping point.

Firstly, consider the trimmed estimate $\hat{N}_k$ itself. Starting from the naïve estimate $\hat{N}_0$, it is expected to decrease towards the ideal estimate $\tilde{N}$ as k increases, provided that trimming is more effective than random sampling. Moreover, according to Result 2, we have $\hat{N}_k > \tilde{N}$ as long as $k < r$. For $k > r$, one can envisage two equilibriums:

1. According to Result 3, ideally, once all the r erroneous records have been removed, we could expect the trimmed estimate to flatten out at the level of the ideal estimate $\tilde{N}$, as k increases.
2. Or, as one gradually exhausts all the effective means, the trimming becomes more or less random at picking out erroneous records. The trimmed estimate would then flatten out at a level higher than $\tilde{N}$, as k increases. How large the bias is depends on the proportion of the erroneous records that remain.

In practice, therefore, one could iteratively trim, adding more records with each step, and monitor $\hat{N}_k$, as k increases, to see if it flattens out at some stage.

Secondly, when considering variance, it is intuitive that k_1, the number of trimmed records confirmed to be non-erroneous, should be as low as possible. Denote by p the probability that a trimmed record is actually erroneous. Let $k_r = r/p$ be the expected number of records, in order to trim the r erroneous records in list A. Then, for any $k < k_r$, the expected number of non-erroneous records is $k(1 - p)$, and homogeneous capture of list B enumeration with probability π implies that the expectation of k_1 is given by

$$E[k_1|k, k < k_r] = k(1 - p)\pi$$

Whereas, for any $k > k_r$, the expected number of non-erroneous records would be $k - r$, so that the corresponding expectation of k_1 is given by

$$E[k_1|k, k > k_r] = (k - r)\pi$$

Thus, k_1 is expected to increase at a rate of $(1 - p)\pi$ as k increases towards k_r, which then changes to π after k becomes larger than k_r. On the one hand, the closer p is to one, or the more effective the trimming is at picking out the erroneous records, the bigger the change. On the other hand, in the case of random trimming or worse, we would have $p \leq r/x$ and $k_r \geq x$. Since it is not possible to trim more than x records in list A, one cannot expect to detect any change in the ratio k_1/k with any such trimming method.

It should be pointed out that, in reality, it is unlikely that the probability p of trimming erroneous records will be a constant of k, i.e. the number of records trimmed. However, the above consideration suggests that, in practice, one could repeat the trimming to successively include more records, and to keep track of the actual k_1, as k increases, for an indication of when to stop. Since it seems natural that the probability p should gradually decrease once the most probable erroneous records have been trimmed, k_1/k may be roughly convex, in which case the stopping point could be where the bend is most acute when plotting $y = k_1$ against $x = k$ (See figure.2.1 for an illustration).

Thirdly, because the way in which the expected value of k_1 changes with k is different before and after $k = k_r$, one can also expect the variance estimate v_k to behave differently before and after k_r, thus providing a third indicator.

The three stopping rules above are all aimed at the same stopping point $k_r = r/p$.

Figure 2.1 provides an illustration.

There are three different settings of (N, n, x, r, p), one for each row of plots. These represent, respectively, a favourable scenario with high capture probability π and reasonably high probability p of trimming erroneous records, an unfavourable scenario with both low π and p, and a scenario between these with low π but reasonably high p.

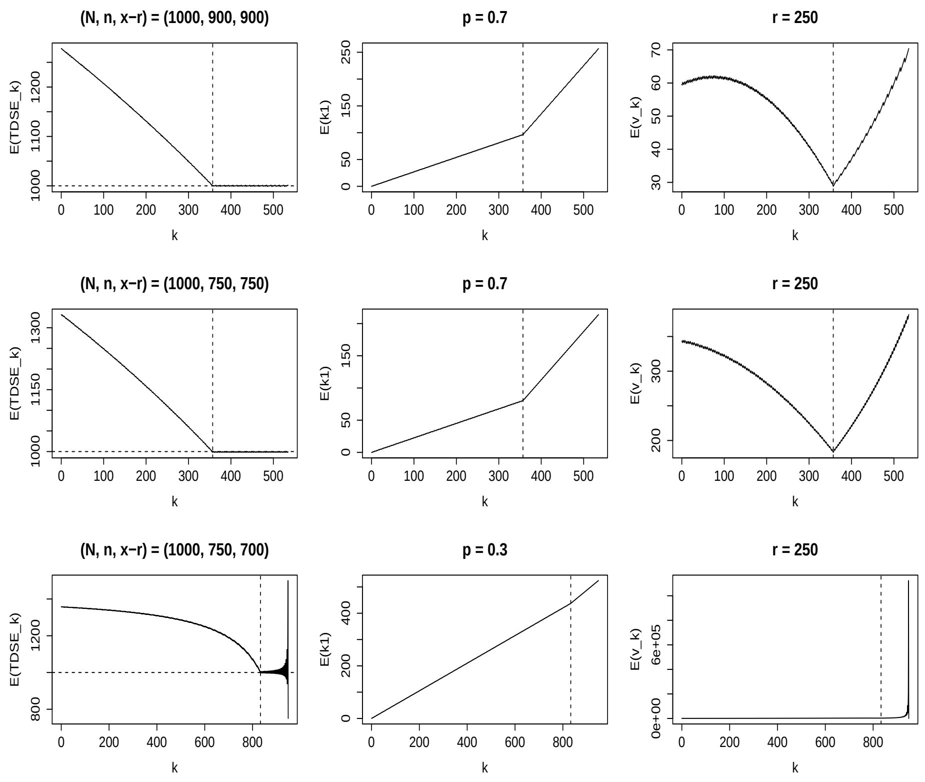

Figure 2.1: Illustration of three stopping rule indicators: left column: $E(\hat{N}_k)$; middle column: $E(k_1)$; right column: $E(v_k)$. Setting (N, n, x, r, p): same for each row.

More explicitly, the target population size is $N = 1000$ in every case. The capture probability of list B is given indirectly as n/N, which is reasonably high at 0.9 in the first setting, and relatively low at 0.75 in the other two. The proportion of erroneous records in list A is given by r/x, which is relatively high at over 20% (i.e. 250/1150) in the first setting, and even higher (i.e. 250/1000 and 250/900) in the other two. The probability p of trimming erroneous records is reasonably high at 0.7 in the first two settings, but rather low at 0.3 in the last one.

It can be seen that all three stopping rules point to the same expected critical point $k_r = r/p$, which is 357 in the first two settings and 833 in the last one. In the first favourable setting, the trimmed DSE becomes unbiased after removing 107 ($= 357 - 250$) extra records compared to the ideal DSE $\tilde{N}$. The standard error (SE) of $\hat{N}_{357}$, on removing all the erroneous records, is $\sqrt{v_{357}} = 5.4$, compared to that of the ideal DSE, i.e. $\sqrt{\tilde{v}} = 3.5$. Still, the loss of efficiency seems a relatively small price to pay compared to the bias of the untrimmed DSE ($\approx \hat{N}_0 - \tilde{N} = 278$).

Similarly in the second scenario with low capture probability π but reasonably effective trimming strategy with probability p. The SE of the trimmed DSE is 13.6 at $k_r = 357$ compared to 10.5 of $\tilde{N}$. Again, a relatively small price to pay against the bias of the untrimmed DSE, which is approximately 332.

In the last unfavourable scenario, the bias of the naïve DSE is 357 to start with. The probability $p = 0.3$ is not much higher than if trimming randomly (at the rate 250/950) in this case. Removing all the erroneous records at such a rate requires on expectation trimming 833 records out of 950 in list A, at which the SE of the trimmed DSE is 50.8 compared to 12.0 of the ideal DSE. Although this may still seem worthwhile in terms of the trade-off between bias and variance, it is unlikely that such a precision is acceptable in practice. However, a poor trimming strategy such as in this scenario still has value in terms of diagnostics and will motivate the search for a more effective trimming strategy.

In this illustration we use expected values. Figure 3.5, in section 3.4, provides an example of similar plots from real data with actual values plotted. In a real life application, the easiest method of identifying the stopping points may simply be observing the behaviour of the estimate with its confidence interval as trimming progresses. In reality, the stopping point does not have to be at the exact point trimming is no longer effective, it can be after this point provided the increase in variance is acceptable.

In summary, the performance of trimmed DSE is above all determined by how effectively the trimming removes the erroneous records. The trimmed DSE can yield good bias-variance trade-off compared to the naïve DSE, even when a considerable number of records are removed from the estimator. Of course, in practice, it may be impossible to remove all the erroneous records by trimming, or one may lack very effective means of trimming. But even then the trimmed DSE can be less biased than the untrimmed one, and it can provide useful sensitivity analysis, because it is easy to compute and interpret.

The methodology presented in this section is used to explore the robustness (or otherwise) of the PECADO system of population estimates in chapter 3. In particular, we note a broader application of the methodology in evaluating the contribution made by each data source contributing to the SPD.

2.4 Exploring impact of heterogeneity in capture rates with respect to list B

2.4.1 Exploring impact on estimates due to heterogeneity in capture rates

To consider the impact of a violation of the *homogeneous capture assumption*, or heterogeneity in the capture rates over specific sub groups, when estimating the size of a specific population cohort, we take as our starting point a simple DSE setup.

We start with a population U of size N. To estimate this population 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 n/N of being caught in list B. Assuming perfect matching we obtain a simple DSE estimator of the population as $\tilde{N} = nx/m$ where m is the size of list $A \cap B$, 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 $n_1/N_1 \neq n_2/N_2$ then a two part (stratified) DSE estimator (equation 2.9) is a more sensible and unbiased estimator for the population U than the simple DSE estimator (equation 2.10).

$$\hat{N} = \hat{N}_1 + \hat{N}_2 = \frac{n_1 x_1}{m_1} + \frac{n_2 x_2}{m_2} \tag{2.9}$$

and noting $n = n_1 + n_2$, $x = x_1 + x_2$ and $m = m_1 + m_2$, we can compare to equation 2.10 where it is assumed $n_1/N_1 = n_2/N_2$

$$\tilde{N} = \frac{nx}{m} = \frac{(n_1 + n_2)(x_1 + x_2)}{(m_1 + m_2)} \tag{2.10}$$

The estimator in equation 2.9 allows for differences in the capture rate for for list B between the two partitions while the estimator in equation 2.10 requires the capture rate to be the same between the two partitions. Now, we can explore the difference between the two estimators, $\tilde{N} - \hat{N}$, to consider the impact of any violation of the *homogeneous capture assumption* for the population U in list B.

We now rewrite equations 2.9 and 2.10 below as 2.11 and 2.12 such that they have the same denominator, mm_1m_2, we also use the fact that $m = m_1 + m_2$ as part of manipulating the numerator to align terms.

$$\begin{aligned}\hat{N} &= \frac{n_1x_1}{m_1} + \frac{n_2x_2}{m_2} \\ &= \frac{mm_2n_1x_1 + mm_1n_2x_2}{mm_1m_2} \\ &= \frac{(m_1+m_2)m_2n_1x_1 + (m_1+m_2)m_1n_2x_2}{mm_1m_2} \\ &= \frac{m_1m_2n_1x_1 + m_2^2n_1x_1 + m_1m_2n_2x_2 + m_1^2n_2x_2}{mm_1m_2} \end{aligned} \tag{2.11}$$

$$\begin{aligned}\tilde{N} &= \frac{(n_1+n_2)(x_1+x_2)}{(m_1+m_2)} \\ &= \frac{n_1x_1 + n_2x_1 + n_1x_2 + n_2x_2}{m} \\ &= \frac{m_1m_2n_1x_1 + m_1m_2n_2x_1 + m_1m_2n_1x_2 + m_1m_2n_2x_2}{mm_1m_2} \end{aligned} \tag{2.12}$$

Taking equations 2.11 and 2.12, and eliminating common terms in the numerator we obtain an equation for $\hat{N} - \tilde{N}$ and simplify below in equation 2.13

$$\begin{aligned}\tilde{N} - \hat{N} &= \frac{m_1m_2n_1x_2 + m_1m_2n_2x_1 - m_2^2n_1x_1 - m_1^2n_2x_2}{mm_1m_2} \\ &= \frac{m_2n_1(m_1x_2 - m_2x_1) - m_1n_2(m_1x_2 - m_2x_1)}{mm_1m_2} \\ &= \frac{(m_2n_1 - m_1n_2)(m_1x_2 - m_2x_1)}{mm_1m_2} \end{aligned} \tag{2.13}$$

We now examine the terms $(m_2n_1 - m_1n_2)$ and $(m_1x_2 - m_2x_1)$ and rewrite them in terms of the coverage rates for list A and list B in the respective population groups as follows

$$\begin{aligned} m_2 n_1 - m_1 n_2 &= \frac{n_1 n_2 x_2}{\hat{N}_2} - \frac{n_1 n_2 x_1}{\hat{N}_1} \\ &= n_1 n_2 \left(\frac{x_2}{\hat{N}_2} - \frac{x_1}{\hat{N}_1} \right) \end{aligned} \tag{2.14}$$

and

$$\begin{aligned} m_1 x_2 - m_2 x_1 &= \frac{x_2 n_1 x_1}{\hat{N}_1} - \frac{x_1 n_2 x_2}{\hat{N}_2} \\ &= x_1 x_2 \left(\frac{n_1}{\hat{N}_1} - \frac{n_2}{\hat{N}_2} \right) \end{aligned} \tag{2.15}$$

and now substituting back into equation 2.13 we get an estimator for the difference between the two part estimator DSE estimator, $\hat{N}$, and the simple DSE estimator, $\tilde{N}$ as $\hat{B}$.

$$\begin{aligned} \hat{B} &= \tilde{N} - \hat{N} \\ &= \left(\frac{n_1}{\hat{N}_1} - \frac{n_2}{\hat{N}_2} \right) \left(\frac{x_2}{\hat{N}_2} - \frac{x_1}{\hat{N}_1} \right) \frac{x_1 x_2 n_1 n_2}{m m_1 m_2} \\ &= \left(\frac{n_1}{\hat{N}_1} - \frac{n_2}{\hat{N}_2} \right) \left(\frac{x_2}{\hat{N}_2} - \frac{x_1}{\hat{N}_1} \right) \frac{\hat{N}_1 \hat{N}_2}{m} \end{aligned} \tag{2.16}$$

If we examine equation 2.16, we clearly see that the sign of $\hat{B}$ is determined by the difference in the coverage rates for lists A and B in the population subgroups U_1 and U_2. We use the term coverage rate as a more general expression of capture rate to facilitate that list A can be any fixed subset of the population. In particular, we can make the following points with regard to the difference between the simple estimator $\tilde{N}$ and the unbiased estimator $\hat{N}$.

- Even if there is a difference in the catch rate for list B between population groups 1 and 2, $(n_2/\hat{N}_2) - (n_1/\hat{N}_1) \neq 0$, the estimators $\tilde{N}$ and $\hat{N}$ will still be equal if the coverage of list A in the different population groups is equal, i.e., $x_1/\hat{N}_1 = x_2/\hat{N}_2$.

This, in effect, would occur if list A complies with *homogeneous capture assumption*, one list still complies with the assumption. Again, Zwane et al. (2004) have a similar conclusion in their paper. However, it also occurs if there is heterogeneity in the capture rates provided the coverage rates for list A in U_1 and U_2 are equal.

- If the direction of the difference in the coverage rates between list A and list B across the population groups, U_1 and U_2, is the same, $\tilde{N}$ will be less than $\hat{N}$. That is if $n_2/\hat{N}_2 > n_1/\hat{N}_1$ and $x_2/\hat{N}_2 > x_1/\hat{N}_1$ or $n_2/\hat{N}_2 < n_1/\hat{N}_1$ and $x_2/\hat{N}_2 < x_1/\hat{N}_1$ then $\tilde{N}$ will be less than $\hat{N}$, the stratified estimator.

- If the direction of the difference in the coverage rates between list A and list B across the population groups, U_1 and U_2, is not the same, $\tilde{N}$ will be greater than $\hat{N}$. That is if $n_2/\hat{N}_2 > n_1/\hat{N}_1$ and $x_2/\hat{N}_2 < x_1/\hat{N}_1$ or $n_2/\hat{N}_2 < n_1/\hat{N}_1$ and $x_2/\hat{N}_2 > x_1/\hat{N}_1$ then $\tilde{N}$ will be greater than $\hat{N}$, the stratified estimator.

- The magnitude of the difference between $\tilde{N}$ and $\hat{N}$ depends on the difference in the coverage rates of list A and list B across the population groups, the greater the difference in the coverage rates the greater the difference in the estimators.

- The magnitude of the difference between $\tilde{N}$ and $\hat{N}$ also depends on how the population is split across the two subgroups; the smaller the difference, $N_1 - N_2$, the greater the difference in the estimators.

- The magnitude of the difference between $\tilde{N}$ and $\hat{N}$ also depends on the size of the general match, m, between lists A and B across the population sub-groups. The larger the general match, the lower the difference in the estimators. This finding relates to that reported by Gerritse et al. (2016) where they conclude that the robustness of the population size estimate is a function of implied coverage. In their context if the implied coverage is high, the match rate will be high and as such the robustness of the estimator will be stronger.

2.4.2 Is the difference due to heterogeneity in capture rates significant?

We consider the general case where heterogeneity in capture rates may occur across h subgroups. We propose to test the null hypothesis $H_0 : D = 0$ where

$$D = \tilde{N} - \hat{N}$$

with

$$\tilde{N} = \frac{nx}{m} = \frac{\sum_h n_h \sum_h x_h}{\sum_h m_h}$$

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

We can write the variance of D as

$$V[D] = V[\tilde{N}] + V[\hat{N}] - 2Cov[\tilde{N}, \hat{N}] \tag{2.17}$$

As both $\tilde{N}$ and $\hat{N}$ are based on DSE methodology, section 2.2.2, we can use equation 2.5 to develop variance estimators for $V[\hat{N}]$, $[\tilde{N}]$.

$$V[\tilde{N}] = \frac{n(n-m)x(x-m)}{m^3} \tag{2.18}$$

$$V[\hat{N}] = \sum_h \left(\frac{n_h(n_h - m_h)x_h(x_h - m_h)}{m_h^3} \right) \tag{2.19}$$

and noting, under the null hypothesis where $m/x = \pi$ that $m_g/x_g = \pi_g = \pi$, that $(x-m)/x = 1 - \pi$ and $m^3 = x^3\pi^3$, we can rewrite $V[\hat{N}]$ as

$$V[\hat{N}] = \sum_h \left(\frac{n_h(n_h - m_h)}{x_h} \frac{(1-\pi)}{\pi^3} \right) \tag{2.20}$$

This then leaves only $Cov[\hat{N}, \tilde{N}]$ to estimate to obtain a variance estimator for $D = \tilde{N} - \hat{N}$.

$$\begin{aligned} Cov[\tilde{N}, \hat{N}] &= Cov\left[\frac{xn}{m}, \sum_h \frac{x_h n_h}{m_h} \right] \\ &= \sum_h Cov\left[\frac{xn}{m}, \frac{x_h n_h}{m_h} \right] \end{aligned} \tag{2.21}$$

Given that x and x_g are fixed, we simply need to find the $Cov[n/m, n_g/m_g]$. n/m and n_g/m_g can now be written in terms of their expectation as follows:

To simplify notation we use μ to designate the expected value.

$$\frac{n}{m} \doteq \frac{\mu_n}{\mu_m} - \frac{\mu_n}{\mu_m^2}(m-\mu_m) + \frac{1}{\mu_m}(n-\mu_n)$$

$$\frac{n_g}{m_g} \doteq \frac{\mu_{n_g}}{\mu_{m_g}} - \frac{\mu_{n_g}}{\mu_{m_g}^2}(m_g-\mu_{m_g}) + \frac{1}{\mu_{m_g}}(n_g-\mu_{n_g}) \tag{2.22}$$

This allows $Cov[n/m, n_g/m_g]$ to be written as (noting the ratio of two expected values is a constant allows the dropping of terms going from line 2 to line 3)

$$\begin{aligned}
&Cov\left[\frac{n}{m}, \frac{n_g}{m_g}\right] \\
&= Cov\left[\frac{\mu_n}{\mu_m} - \frac{\mu_n}{\mu_m^2}(m-\mu_m) + \frac{1}{\mu_m}(n-\mu_n), \frac{\mu_{n_g}}{\mu_{m_g}} - \frac{\mu_{n_g}}{\mu_{m_g}^2}(m_g-\mu_{m_g}) + \frac{1}{\mu_{m_g}}(n_g-\mu_{n_g})\right] \\
&= Cov\left[-\frac{\mu_n}{\mu_m^2}(m-\mu_m) + \frac{1}{\mu_m}(n-\mu_n), -\frac{\mu_{n_g}}{\mu_{m_g}^2}(m_g-\mu_{m_g}) + \frac{1}{\mu_{m_g}}(n_g-\mu_{n_g})\right] \\
&= Cov\left[-\frac{\mu_n}{\mu_m^2}(m-\mu_m), -\frac{\mu_{n_g}}{\mu_{m_g}^2}(m_g-\mu_{m_g})\right] + Cov\left[-\frac{\mu_n}{\mu_m^2}(m-\mu_m), \frac{1}{\mu_{m_g}}(n_g-\mu_{n_g})\right] + \\
&\quad Cov\left[\frac{1}{\mu_m}(n-\mu_n), -\frac{\mu_{n_g}}{\mu_{m_g}^2}(m_g-\mu_{m_g})\right] + Cov\left[\frac{1}{\mu_m}(n-\mu_n), \frac{1}{\mu_{m_g}}(n_g-\mu_{n_g})\right] \\
&= \left(\frac{-\mu_n}{\mu_m^2}\right)\left(\frac{-\mu_{n_g}}{\mu_{m_g}^2}\right) Cov[m, m_g] + \left(\frac{-\mu_{n_g}}{\mu_{m_g}^2}\right)\frac{1}{\mu_m} Cov[n, m_g] + \\
&\quad \left(\frac{-\mu_n}{\mu_m^2}\right)\frac{1}{\mu_{m_g}} Cov[n_g, m] + \frac{1}{\mu_m}\frac{1}{\mu_{m_g}} Cov[n, n_g]
\end{aligned} \tag{2.23}$$

We now examine each of the terms $Cov[n, n_g]$, $Cov[n, m_g]$, $Cov[n_g, m]$ and $Cov[m, m_g]$ under the assumption of independent capture, the event that one unit in the population is caught in list B is independent of any of the event of any other unit being caught in list B as follows

$$\begin{aligned}
Cov[n, n_g] &= Cov[n_g + (n - n_g), n_g] \\
&= Cov[n_g, n_g] + Cov[(n - n_g), n_g] \\
&= V[n_g]
\end{aligned}$$

$$\begin{aligned}
Cov[n, m_g] &= Cov[m_g + (n - m_g), m_g] \\
&= Cov[m_g, m_g] + Cov[(n - m_g), m_g] \\
&= V[m_g]
\end{aligned}$$

$$\begin{aligned}
Cov[n_g, m] &= Cov[m_g + (n_g - m_g), m_g + (m - m_g)] \\
&= Cov[m_g, m_g] + Cov[m_g, (m - m_g)] + \\
&\qquad Cov[(n_g - m_g), m_g] + Cov[(n_g - m_g), (m - m_g)] \\
&= V[m_g]
\end{aligned}$$

$$\begin{aligned}
Cov[m, m_g] &= Cov[m_g + (m - m_g), m_g] \\
&= Cov[m_g, m_g] + Cov[(m - m_g), m_g] \\
&= V[m_g]
\end{aligned} \tag{2.24}$$

and using the binomial distribution under the null hypothesis, express the covariance terms as

$$\begin{aligned}
Cov[n, n_g] &= V[n_g] = N_g \pi(1 - \pi) \\
Cov[n, m_g] &= V[m_g] = x_g \pi(1 - \pi) \\
Cov[n_g, m] &= V[m_g] = x_g \pi(1 - \pi) \\
Cov[m, m_g] &= V[m_g] = x_g \pi(1 - \pi)
\end{aligned} \tag{2.25}$$

We now look again at the expression $Cov[xn/m, x_g n_g/m_g]$ and write

$$\begin{aligned}
&Cov\left[\frac{xn}{m},\frac{x_gn_g}{m_g}\right]\\
&\quad= xx_gCov\left[\frac{n}{m},\frac{n_g}{m_g}\right]\\
&\quad= xx_g\frac{1}{\mu_m}\frac{1}{\mu_{m_g}}V[n_g] + xx_g\left(\frac{-\mu_{n_g}}{\mu_{m_g^2}}\right)\frac{1}{\mu_m}V[m_g]+\\
&\qquad xx_g\left(\frac{-\mu_n}{\mu_m^2}\right)\frac{1}{\mu_{m_g}}V[m_g] + xx_g\left(\frac{-\mu_n}{\mu_m^2}\right)\left(\frac{-\mu_{n_g}}{\mu_{m_g}^2}\right)V[m_g]\\
&\quad= xx_g\frac{1}{\mu_m\mu_{m_g}}V[n_g] + xx_g\left(\frac{\mu_n\mu_{n_g}}{\mu_m^2\mu_{m_g}^2} - \frac{\mu_n}{\mu_m^2\mu_{m_g}} - \frac{\mu_{n_g}}{\mu_{m_g}^2\mu_m}\right)V[m_g] \qquad (2.26)
\end{aligned}$$

Under the null hypothesis, $\mu_m = x\pi$ and $\mu_{m_g} = x_g\pi$ which in turn now allows us to continue with

$$\begin{aligned}
&Cov\left[\frac{xn}{m},\frac{x_gn_g}{m_g}\right]\\
&\quad= x_gxx_g\pi(1-\pi)\left[\frac{\mu_n\mu_{n_g}}{x^2\pi^2x_g^2\pi^2} - \frac{\mu_n}{x^2\pi^2x_g\pi} - \frac{\mu_{n_g}}{x_g^2\pi^2x\pi}\right]+\\
&\qquad xx_g\frac{1}{x\pi x_g\pi}N_g\pi(1-\pi) \qquad (2.27)
\end{aligned}$$

and with $\hat{\pi} = m/x$ and $\hat{N}_g = n_g/\hat{\pi} \implies \hat{\mu}_{n_g} = \hat{N}_g\hat{\pi} = n_g \implies \hat{\mu}_n = n$ we can now write

$$\begin{aligned}
&\widehat{Cov}\left[\frac{xn}{m},\frac{x_gn_g}{m_g}\right]\\
&\quad= x_gxx_g(1-\hat{\pi})\left[\frac{nn_g}{x^2\hat{\pi}^2x_g^2\hat{\pi}} - \frac{n}{x^2\hat{\pi}^2x_g} - \frac{n_g}{x_g^2\hat{\pi}^2x}\right]+\\
&\qquad xx_g\frac{1}{x\hat{\pi}x_g\hat{\pi}}\frac{n_g}{\hat{\pi}}\hat{\pi}(1-\hat{\pi})\\
&\quad= \frac{(1-\hat{\pi})}{\hat{\pi}^2}\left[\frac{nn_g}{x\hat{\pi}} - \frac{nx_g}{x} - n_g + n_g\right]\\
&\quad= \frac{(1-\hat{\pi})}{\hat{\pi}^2}\left[\frac{nn_g}{m} - \frac{nx_g}{x}\right] \qquad (2.28)
\end{aligned}$$

and now summing over strata, h, we obtain an expression for $\widehat{Cov}[\tilde{N},\hat{N}]$ as follows

$$\begin{aligned}\widehat{Cov}[\tilde{N},\hat{N}] &= \frac{(1-\hat{\pi})}{\hat{\pi}^2}\left[\frac{n\sum_h n_h}{m} - \frac{n\sum_h x_h}{x}\right]\\ &= \frac{(1-\hat{\pi})}{\hat{\pi}^2}\left[\frac{n^2}{m} - n\right]\end{aligned} \tag{2.29}$$

We can now consider a test statistics $Z = \hat{D}/\sqrt{V[\hat{D}]}$ as N(0,1) in testing the null hypothesis $H_0 : \hat{D} = 0$ where

$$\hat{D} = \tilde{N} - \hat{N}$$

$$V[\hat{D}] = V[\tilde{N}] + V[\hat{N}] - 2Cov[\hat{N},\tilde{N}]$$

To demonstrate the test, we consider a population divided into 3 strata with list sizes, (x_h, n_h, m_h), for each stratum as (50,40,20), (100,90,30) and (150,120,90). This data provides population size estimators $\tilde{N} = 536$ and $\hat{N} = 600$ and values for $D = 64$, $V[D] = 12$ and the test statistic $Z = 18$ indicating significant bias due to heterogeneity in capture rates across subgroups if heterogeneity in capture rates is ignored.

2.5 Concluding Remarks

Development of the PECADO toolkit in this way allows for the application of DSE methods in many more situations. A particular innovation in how the DSE is derived in section 2.2.2 is that list A is considered fixed. This has significant advantages as will be seen later in chapter 5 when considering extensions of this method to estimate population flows.

In chapter 3, the proposed new system of population estimates relies on the development of the DSE methods in this way as list A is compiled from administrative data sources leaving only list B being required to satisfy the *homogeneous capture assumption*. A further consideration of this assumption in section 2.4 allows for heterogeneity in the capture rates across subgroups, provided the coverage rates across the same subgroups for list A are constant. This further consideration strengthens the argument for the choice of data source to use as list B in the PECADO system.

We now consider the *homogeneous capture assumption* in this more relaxed form as part of the PECADO toolkit, that is, we will also consider the case of heterogeneity between subgroups as valid under this assumption provided the coverage rates for the same subgroups in list A are equal. Section 2.4.2 provides a significance test with respect to heterogeneity in capture rates with respect to list B.

In addition the deployment of TDSE to hunt for potential erroneous records provides added reassurance that the potential for bias in the final population estimates due to erroneous records being present in list A is minimised.

In summary, the PECADO toolkit as developed in this chapter, can also be in various diagnostic ways to evaluate underlying assumptions as well as provide reassurance around the robustness of the final estimate.

In chapter 4 we use the same DSE methods to evaluate the possibility of undercoverage in the Census counts for 2016 while in chapter 5 we incorporate the same DSE methods to extend the capability of the PECADO system to also estimate gross population flows (inflows and outflows).

If the traditional Petersen model, described in section 2.2.1, was used it would be very difficult to justify that the underlying assumptions held for the use cases in each of chapters 3, 4 and 5. The PECADO toolkit has not considered the Chapman formula (Chapman, 1951) for correcting for bias when cell sizes are very small. The Chapman formula contains a minor adaption to the traditional Petersen model formula that, in particular, can cater for a null match between the two lists when cell sizes are small. This thesis will primarily deal with much larger cell sizes associated with State level estimates and as such the Chapman adaptation is not considered to have an impact on the estimates. However, the Chapman adaptation would become a bigger consideration if considering the direct compilation of population estimates in much greater geographical detail using DSE methods.

Zhang, in a recent paper (Zhang, 2019), also gives further consideration to these DSE methods and their underlying assumptions.

Chapter 3

PECADO - a robust system of population estimates

3.1 Introduction

In this chapter we explore the system of population estimates proposed in section 1.3.

To recap, the system of population estimates can be summarised as follows:

We built the PAR, an SPD, using a 'Signs Of Life' (SoL) approach to act as our list A. We then use the Driver Licence Dataset (DLD) to act as our list B in a capture recapture system to adjust for undercount in the PAR. The SoL approach, in theory, eliminates overcoverage as an issue.

In developing this system we have made three key assumptions which we will use to explore the robustness of the system of population estimates.

- No erroneous records: There are no erroneous records in either list A or list B
- No linkage error: There is no linkage error between list A and list B
- Homogeneous capture: Each unit in the population has an equal probability of being included in list B. Where there is heterogeneity in capture rates between sub groups the coverage rates in list A are assumed to be equal.

Blocking is undertaken by gender, year of age and nationality grouping. The attributes for the blocking variables always come from the same underlying master register to ensure no domain incoherence between lists. This ensures no inflation of population estimates due to inconsistencies between lists in the blocking variables gender, age and nationality grouping. Blocking facilitates the homogeneous capture assumption in that if there are differences in capture rates, they will more than likely occur across blocks rather than within blocks. Blocking also facilitates the compilation of the population by those same variables.

The use of deterministic matching with high quality identification numbers ensures negligible or no linkage error. Deterministic matching typically refers to where an exact match between fields identifies whether units are the same. These fields can be identifiers with the explicit purpose of uniquely identifying units. In Ireland the PPSN is used to identify persons on different public administration systems.

In theory, the system is set up so that the population estimates are robust. However, in practice, there may be weaknesses in the underlying assumptions. The purpose of this chapter is to examine how robust the system of population estimates is, given the underlying data sources.

Reference year 2011 is selected as a particular year to examine as population estimates from the Census are also available for this year.

In the next section, section 3.2, we present the population estimates and consider the availability of the underlying data sources. Section 3.3 considers the homogeneous capture assumption. Then section 3.4 considers the presence of erroneous records and presents the Trimmed Dual System Estimation (TDSE) tookit that allows statisticians hunt for overcoverage. Section 3.5 will then consider how dependent the system of population estimates is on each underlying data source.

Sections 3.6 and 3.7 conclude the chapter with a summary of our findings and insights with respect to the proposed system of population estimates.

3.2 A system of population estimates

3.2.1 Availability of underlying data sources

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 examine the dependence of this system on individual data sources in more detail later, in section 3.5.

Table 3.1 provides a summary of the availability of the different data sources by calendar year.

	Year					
	2011	2012	2013	2014	2015	2016
PAR - List A data sources						
Child Benefit (CB)	Y	Y	Y	Y	Y	Y
Early Childhood Care (ECCE)	N	Y	N	N	N	N
Primary School Pupils (POD)	N	N	N	N	N	N
Post Primary Pupils (PPP)	Y	Y	Y	Y	N	N
Higher Education Enrolments (HEA)	Y	Y	Y	Y	Y	N
Further Education Awards (FET)	Y	Y	Y	Y	Y	Y
Employer Employee Tax Returns (P35)	Y	Y	Y	Y	Y	Y
Income Tax Returns (self-employed) (IT)	Y	Y	Y	Y	Y	N
Social Welfare (SW)	Y	Y	Y	Y	Y	Y
Public Health Benefits (PCRS)	N	N	Y	Y	Y	Y
State pension (SP)	Y	Y	Y	Y	Y	Y
PAR - List B data sources						
Driver Licence Dataset (*DLD*)	Y	Y	Y	Y	Y	Y
Quarterly National Household Survey (*QNHS*)	Y	Y	Y	Y	Y	Y

Table 3.1: Availability of Data Sources by year. ECCE available only for 2012 but expected to be available for future years. POD to be available from reference year 2017. $QNHS$ is further described in section 3.3.2

Figure 3.1 illustrates the coverage of each source for the chosen reference year with respect to the PAR.

In looking at the school age and pre school age part of the population pyramid in figure 3.1, we see that the counts are very much dependent on the Child Benefit (CB) data source. The Post Primary Pupils (PPP) data source only covers part of this population group. We also note a small but significant gap between PPP and CB.

In looking at the working age population, we see that the main sources of activity are P35 (a list of all employees and those in receipt of occupational pensions for that year, as provided by employers and pension administrators) and SW (those in receipt of some type of social benefit from the state). Another significant source in this age category for females is CB (those parents in receipt of a child benefit payment, which is typically paid to the mother of the child).

In considering those over 65 years of age or in retirement the primary data source is SP which practically equates to the full SPD count. It should be noted here that as the actual State Pension payments data was unavailable, an indicator based on activity in administrative records is used to create a proxy data source. Those in receipt of occupational pensions on the P35 data source also have high coverage (this is significantly higher for males than females). One source not included for this year, but available in later years, is a data source related to public health care (PCRS). The PCRS data source can be expected to have high coverage in the older age groups due to the rules of the system and the demand for health care in this age category.

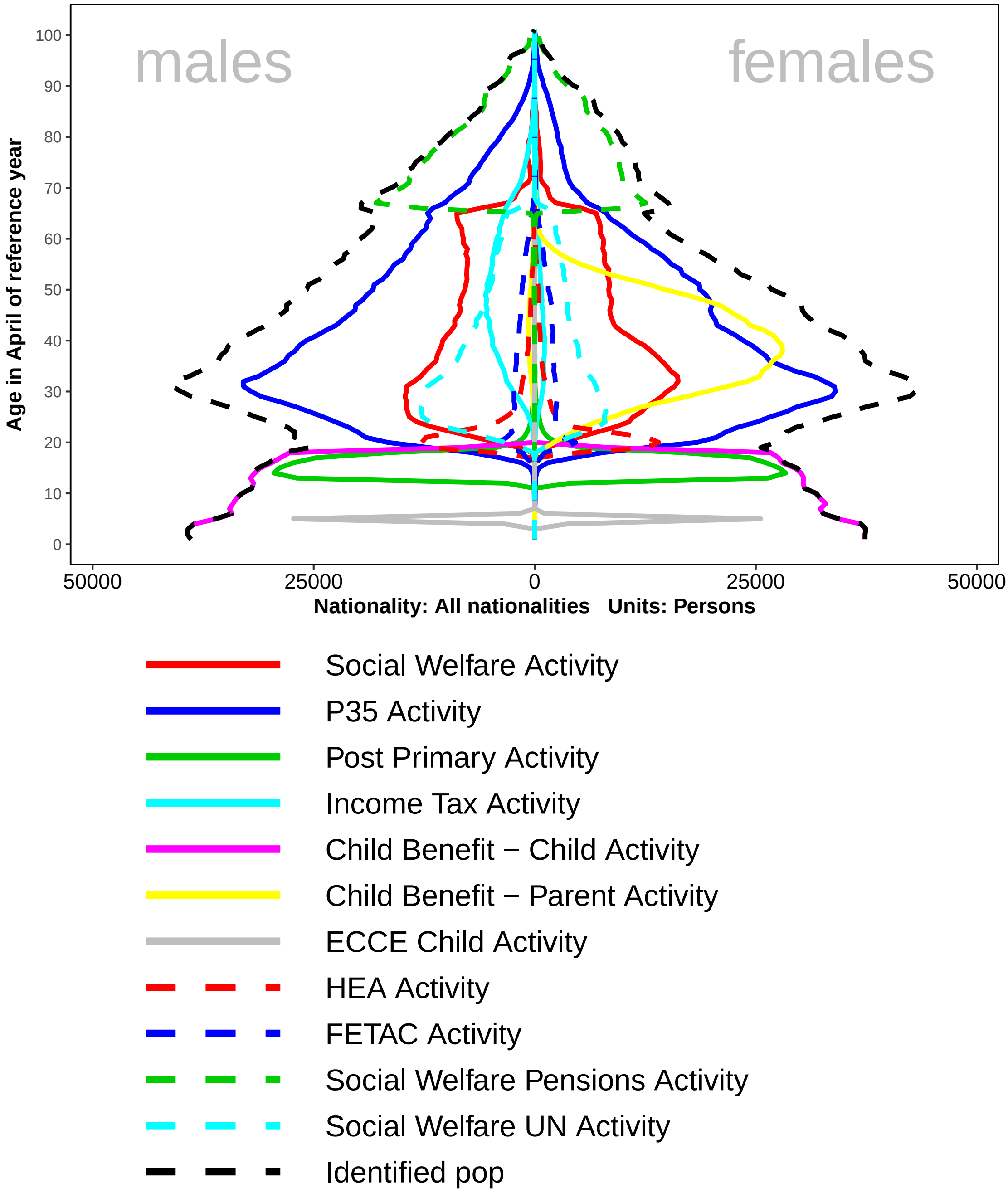

Figure 3.1: Administrative Data Source Coverage with respect to PAR, 2011

Figure 3.2 shows the proportion of driving licence holders identified on the PAR. Note the actual proportion is higher because only those that have renewed or applied for a licence in recent years will have been required to provide a PIN. A driving licence is typically valid for 10 years. A clear difference can be seen between nationality groupings and their propensity to hold an Irish licence. According to the rules for driving in Ireland, UK and EU licence holders may not have a strong motivation to hold an Irish licence as driving licences of these nationalities are recognised in the State for a period of time.

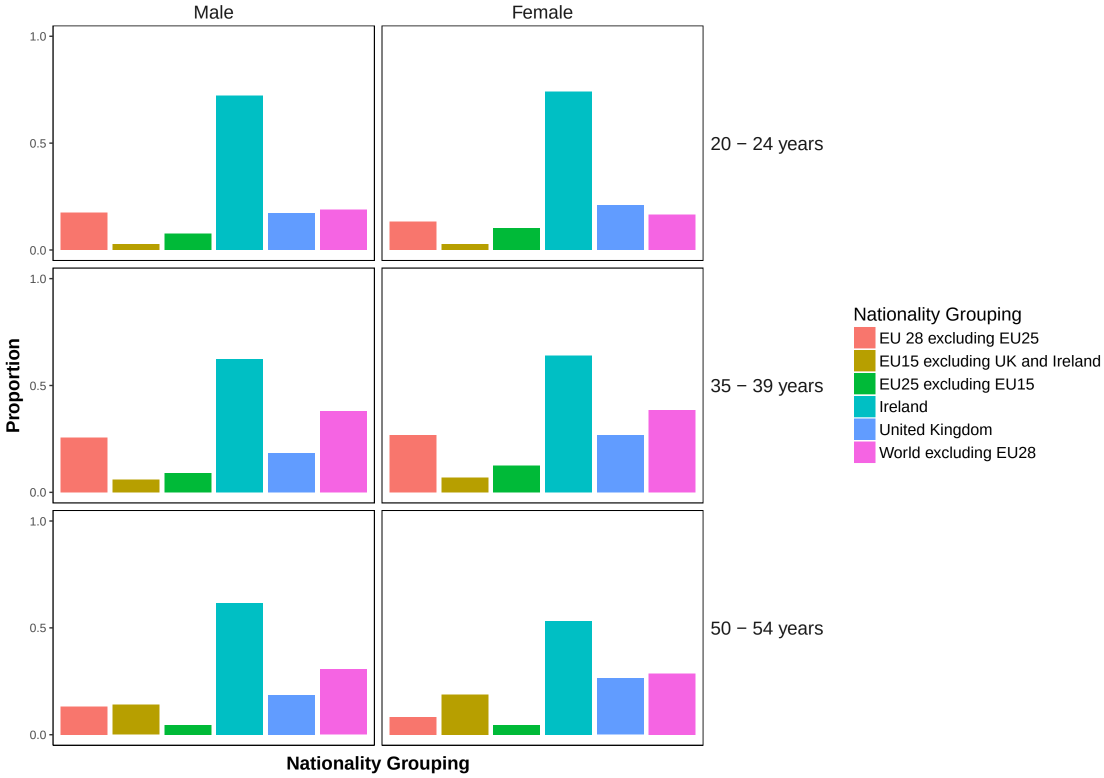

Figure 3.2: Proportion of identified Irish driver licence holders on *PAR* by nationality group, selected age group and sex, 2011

Driving licences originating from outside the EU do not have the same recognition as EU driving licences. This analysis provides justification for blocking by nationality group, age and sex. Blocking or post-stratification is a standard method in Census population size adjustment which helps to account for the heterogeneous capture in the population. Blocking also provides for enhanced Census-like estimates by nationality grouping, age and gender.

As per our analysis above, blocking by age, sex and nationality group will eliminate the requirement for the homogeneous capture assumption to hold between blocks. The homogeneous capture assumption for list B is only required to hold within blocks. Furthermore, we again note from section 2.4 that list B will allow for a difference in the capture rate between two sub-groups without introducing bias to the estimate, provided the coverage rates for those two subgroups in list A are equal. This argument facilitates the consideration of both drivers and non-drivers in the DSE setup without introducing bias, provided we assume no difference between drivers and non drivers with respect to their propensity to interact with public services used in compiling list A. We examine the strength of this assumption later in section 3.3.

3.2.2 Population estimates

We now adjust the SPD counts using the DSE methodology outlined in section 2.2.2 to obtain a first set of population estimates and present them in figure 3.3. The figure presents the population estimates along with the SPD and *DLD* counts using a population pyramid with males to the left, females to the right and age on the y axis. The Census 2011 population counts are also shown to enable an initial evaluation of the estimates. The population estimates also have a 95% Confidence Interval plotted using dotted lines but the relative size of the confidence intervals is such that it is almost impossible to make them out on the plot.

In comparing the SPD counts to the population estimates we make the following observations

- Coverage of the SPD is high
- Significant coverage deficits are obvious just prior to the retirement age of 65, with the deficit being larger for females.
- SPD coverage for those in retirement age also looks to be lower for females than males
- No adjustment is made to the SPD count for children - children don't hold driver licence and as such the *DLD* dataset cannot be used to adjust for undercoverage in these groups
- The confidence intervals for the population estimates are relatively very small. This is due to the high coverage rate of the SPD.

In comparing the Population estimates with the Census 2011 estimates, the following observations can be made:

- For those under 18 the population estimates are close to the Census estimates, albeit slightly higher. This part of the population is not subject to high migration flows.
- The population estimates are significantly higher than the Census estimates for the age category 25 to 40 years. This part of the population may be subject to high migration flows.
- The estimation methodology looks to do a very good job in adjusting the SPD counts from 40 years up, especially where significant coverage deficits have been found just below the retirement age. This part of the population is not subject to high migration flows.

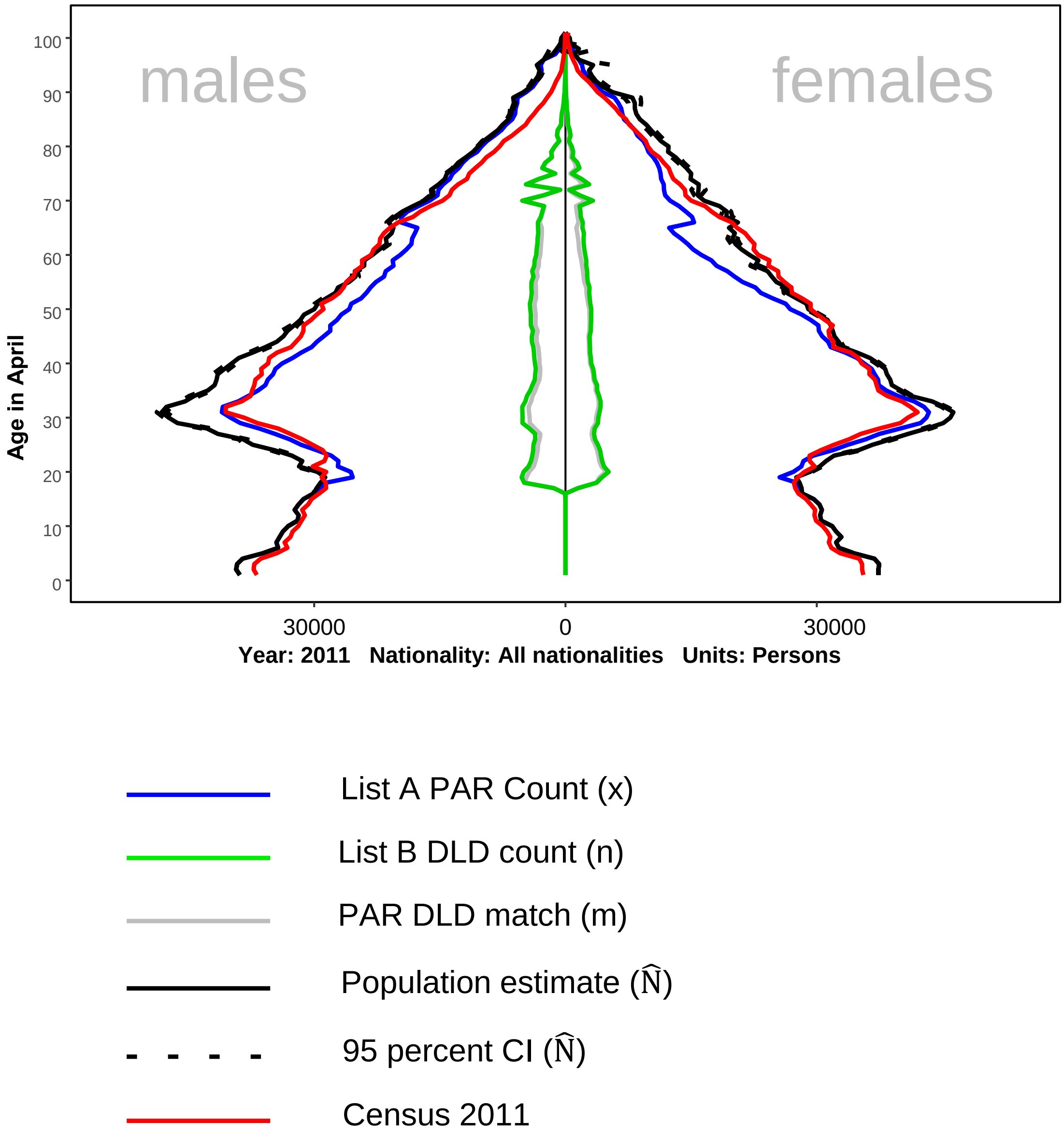

Figure 3.3: Preliminary population estimates by sex and single year of age, All nationalities, 2011

- The population estimates are significantly higher than the Census estimates for males aged over 75.
- The population estimates show a higher number of males than females aged over 75. This age group is not subject to high migration flows.

On the plus side, the system seems to do a good job at adjusting for the significant deficits in SPD coverage in the pre-retirement age groups (retirement age at approximately 65 years). However, areas of concern that warrant further investigation are the significant differences between Census counts for the age groups 25 - 45 years of age and over 75 years of age.

Some observations from figure 3.1 may raise suspicions about erroneous records being included in the *PAR* count. However, under this system as it stands it has no list B to evaluate undercoverage for those persons below the legal age at which you can hold a driver licence. The next section considers another data source that can be used or substituted in as a list B in the proposed system, this second source will also help to validate the use of *DLD* as a suitable list B.

3.3 Evaluation of Driver Licence Dataset (*DLD*) as list B under the homogeneous capture assumption

3.3.1 Evaluation strategy

The *DLD* dataset has already been described in section 1.3.2.2.

The simplest way to evaluate this assumption is to:

- identify an alternative data source that can be used as list B
- 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 then both data sources used as list B can be considered. This similarity will then indicate that both data sources satisfy the homogeneous capture assumption or that both data sources violate the assumption in such a way that the two sets of population estimates are consistent.

If the two sets of population estimates are not consistent, this leads to the conclusion that one or more of the two data sources used as list B violate the homogeneous capture assumption.

This strategy can be deployed within the DSE framework presented in chapter 2 to validate the *homogeneous capture* assumption without resorting to more complicated Triple System Estimation (TSE) type models. Descriptions of triple system estimation models are readily available in textbooks such as Bishop et al. (1975) and Baffour et al. (2013) provides a nice overview of how TSE models with the addition of a third list, list C, can be used to estimate dependence between lists. This approach can also be used to explore the homogeneous capture assumption. If the assumption of *homogeneous capture* in the population with respect to list B holds and the event of a person being captured in list B is independent of any other individual in list B being also captured then the interaction term for lists A and B will not be significant. We also note the typical approach to using TSE involves 3 lists that can capture units of the population and then models those capture rates for each person in the population for those lists and the dependence of capture between each pair of lists, while in the DSE framework presented earlier, list A can be considered as any fixed list from the population (i.e., there is no need to consider capture rates with respect to list A, only the size).

3.3.2 An alternative list B - Quarterly National Household Survey ($QNHS$)

We now consider the Quarterly National Household Survey ($QNHS$) undertaken by CSO, Ireland as an alternative list B.

The $QNHS$ is the name of a survey undertaken by the CSO with the primary purpose of measuring the unemployment rate up until the second quarter 2017. It was replaced by a new quarterly survey with similar characteristics in the third quarter 2017 called the Labour Force Survey (LFS).

The $QNHS$ survey design has the following characteristics

- Issued sample size of 25,000 households per quarter
- Two stage design, where each household has equal probability of being selected
- Achieved sample size of approximately 15,000 households per quarter
- Survey quarter has 5 waves. Each wave stays in for 5 quarters and each quarter a wave is swapped with a new wave (a rotating panel).
- Survey does not include those persons that do not live in a fixed household (i.e. those living in Institutions or having no fixed abode)

As each household is designed to have equal probability of being selected, and given that almost every person in the population lives in a fixed household, each person is also considered to have equal probability of being selected in the sample.

The url, http://www.cso.ie/en/qnhs/qnhsmethodology/ , (accessed on 15th August 2017) has more information on the methodology underpinning the $QNHS$.

The $QNHS$ sample data is further processed to enable it to be used as a list B in compiling population estimates using DSE. The $QNHS$ sample is first matched with the PIN master file in a deterministic way using first name, surname and date of birth as the match keys. Only those records with a unique match against the master file are retained. In theory, it is possible that more than one person may have the same surname, first name and date of birth, but it is highly unlikely. In the likelihood of such an event (identified on the PIN master file), these records are omitted from the list. This results in a list B with approximately 60,000 persons (just over 1% of the population) that are recorded as being usually resident in the State at some stage in the calendar year.

In addition, two further assumptions are made in the preparation of list B. First, it is assumed that non response in the $QNHS$ can be considered as missing at random. [1] Second, it is also assumed that those records from the $QNHS$ dataset that did not match to the master file can also be considered as missing at random. These assumptions are required to ensure that the *homogeneous capture* assumption is valid. Where there is a violation of these assumptions, it may be possible to deploy an appropriate weighting schema to correct counts (list A and AB corresponding to the match) within strata, provided sufficient information is available. In our application, we rely on the two additional assumptions holding within strata (age, sex and nationality grouping).

This list B can now be matched with list A using a high quality PIN. We assume no linkage error. We will label this list B as the $QNHS$ list.

In the next section we compile population estimates using $QNHS$ as list B and compare with those estimates that have used DLD as list B.

3.3.3 DLD V $QNHS$ - a list B comparison

Figure 3.4 presents a comparison of two sets of populations estimates where one set has been compiled using DLD as list B (denoted with blue) and the other set has been compiled using $QNHS$ as list B (denoted in green).

The confidence intervals for the population estimates compiled with $QNHS$ have been estimated as if each capture in list B is independent. In reality this is not the case. The two stage design of the survey and the fact that, if one person in a house is captured for the survey, all individuals in that house will be captured, violates the assumption of independent capture for individuals in estimating the variance as in equation (2.5). The

[1] Internal CSO reports show that response rates differ slightly according to certain household characteristics (urban/rural, dwelling type, number of persons in household etc.) and personal characteristics of head of household (gender, marital status, nationality, PES - Principal Economic Status, etc.). This analysis was based on linking Survey data with Census data.

design effect, as described in Kish (1995), for the $QNHS$ when measuring unemployment rates is estimated in the region of 1.5 to 2. Therefore, the confidence intervals in practice will be wider than those shown for the $QNHS$ based estimates which have been calculated assuming a design effect of 1. If we assume the design effect also carries over to our population estimates then, in theory, the confidence intervals should be adjusted by a factor equal to the square root of the design effect, so if we had a design effect of 2 then the confidence intervals should be adjusted by a factor of $\sqrt{2}$.

In comparing the two sets of figures there is very little difference across age group and gender. The largest differences occur where the population estimates peak for both males and females at age 30 (excluding an odd peak at age 20). However care needs to be taken in determining how significant this difference is, as the estimates of the confidence intervals for the $QNHS$ based estimates are under estimated and a number of assumptions have also been made in asserting that the $QNHS$ list B has been compiled under the homogeneous capture assumption.

If this difference is considered significant, there are a number of possible explanations for this difference.

One possible explanation for this difference may be that the DLD contains some erroneous records, i.e., there may be persons renewing their driving licence who reside outside the State, even though there is a requirement for a person to provide significant evidence that they reside in the State before renewing their driver licence. In statistical terms, this can be expressed as $U_{DLD} > U_{QNHS}$ where U_{DLD} is the hypothetical Driver Licence population with some erroneous records from those living abroad and U_{QNHS} is the hypothetical usually resident population for those usually resident in the State from which the $QNHS$ sample is drawn. This may be an arguable explanation as there is a cost to letting a drivers licence lapse - a person may have to resit their driving test. However, in practice the burden of obtaining evidence (ie., a utility bill or bank statement with your address) is considered a significant deterrent.

However, this argument could also be reversed to give a second explanation. It maybe that the reason $U_{DLD} > U_{QNHS}$ is related to certain groups of the population, say certain cohorts of young males, who are difficult or impossible to reach. These groups may not be properly represented in the $QNHS$ sample, thus leading to an underestimate in the population. This same group would still however be represented properly in the DLD as they require a valid driver licence to drive in the State.

The choice of DLD as list B over $QNHS$ is preferable unless there is sufficient evidence to dismiss the DLD based estimates. The reason for this preference is that the DLD sample is significantly larger and therefore provides more precise estimates.

The one advantage to basing list B on $QNHS$ rather than DLD is that $QNHS$ also covers the pre driving age categories, that is those age categories under 18. In looking

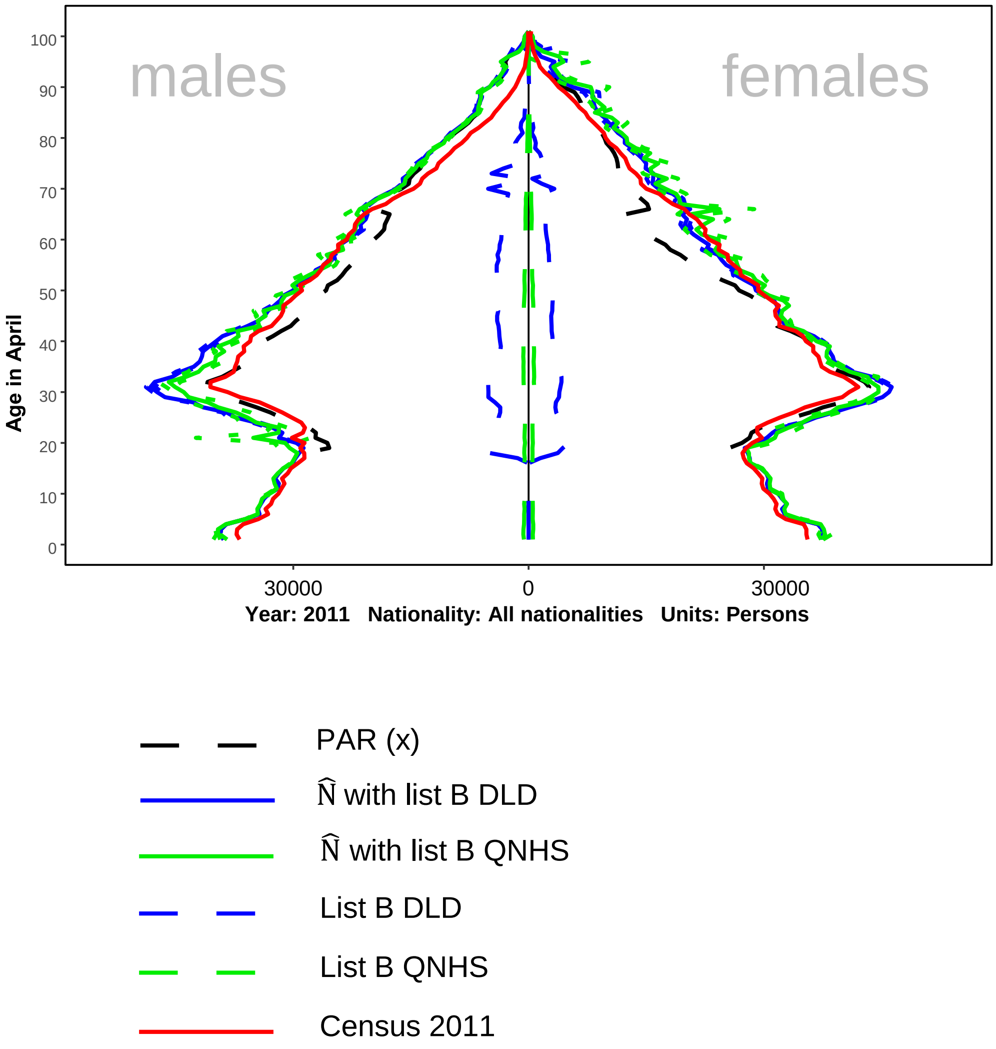

Figure 3.4: Comparison of population estimates using two different data sources as list B, All nationalities, 2011. The first data source used as list B is based on the Quarterly national Household Survey ($QNHS$) while the second is based on the Driver Licence Dataset (DLD). 95% confidence intervals are shown with dots around population estimates.

at figure 3.4 we see almost no adjustment to PAR counts for $QNHS$ based population estimates indicating that there is no undercount in the under 18 years age groups. In fact, the difference between the Census counts and the DSE population estimates suggests there may be overcoverage in the PAR counts (or undercoverage in the Census) for this age group.

3.4 TDSE illustrated

We apply the TDSE, described in section 2.3.2, to our system of population estimates to illustrate an application of the method before applying it in a more strategic manner to evaluate underlying data sources in the PAR.

We trim list A and list AB where list AB is the list of matched units of size m between list A and list B. The criteria for selecting the k records to be trimmed is based on subjectively identifying those records that are most likely to contain erroneous records. In this example, the trimming method removes records for persons in list A in a number of steps where the SoL is based solely on an employment record with earnings less than a specified amount in EUR. The P35 data source also contains information on earnings. So, after finding the base estimate at $\hat{N}_0$ with no trimming, step 1 requires removing records for persons with only an employment record with pay less than €1K, step 2 removes records for persons with only an employment record with pay less that €2K, and so on. We note again that if those persons have another record on another data source indicating SoL, they are not removed.

On examining the TDSE in year 2011 for different post-strata (by age, sex, nationality group) we see that it can behave differently in different post-strata. Fig 3.5 presents 3 different situations (one per row) with respect to the stopping rules described in Section 2.3.3. In the first case (Row 1), the population group relates to males aged 32 with a nationality from the most recent EU countries, referred to as EUnew, and $\hat{N}_k$ shows a distinctive fall before a general levelling off. In the second case (Row 2), the population group relates to males aged 56 years of Irish nationality, and $\hat{N}_k$ appears to be generally level with a possible small general decline over the trimming. In the last case (Row 3), the population group relates to females aged 28 years of Irish nationality, and $\hat{N}_k$ starts generally level before appearing to rise slowly.

More explicitly, the first stopping rule looks to see if $\hat{N}_k$ flattens out at some point, indicating that the trimming method has reached an equilibrium. In considering the 3 cases, the first case looks to have a point $k_r \approx 520$ where $\hat{N}_k$ appears to flatten out at 6460. The second case has no such point while the third case appears to have a point $k_r \approx 700$ where $\hat{N}_k$ starts to rise, indicating that the trimming method is removing fewer erroneous records than would be the case if it were removing the records at random.

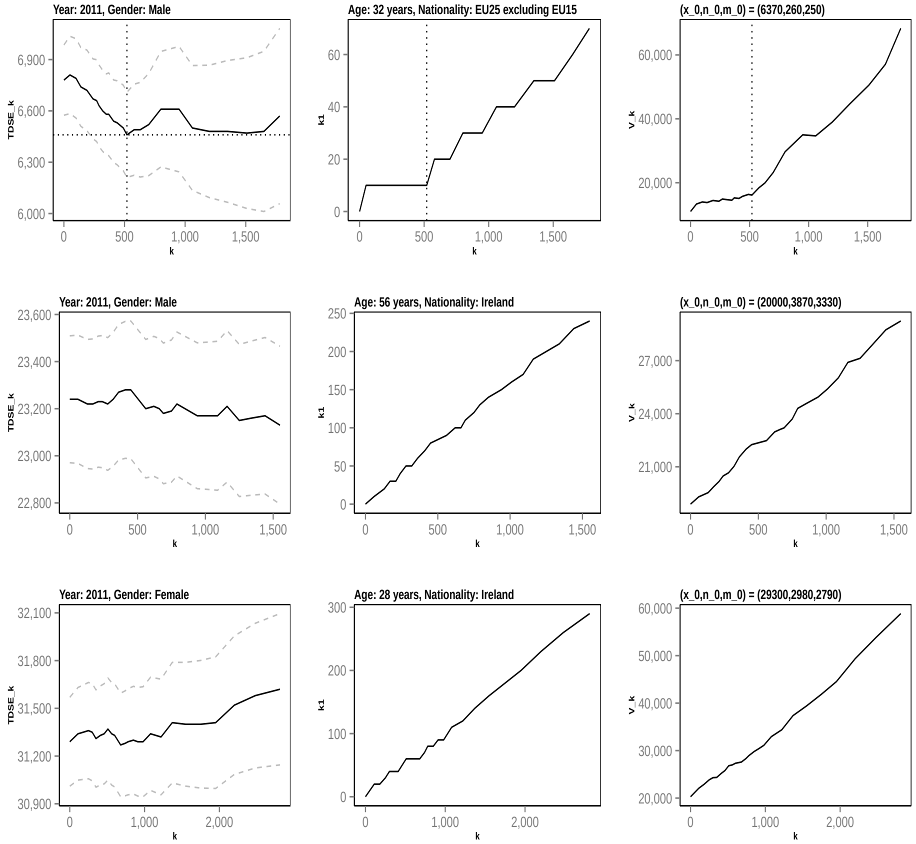

Figure 3.5: Illustration of TDSE in year 2011. Left column: TDSE $\hat{N}_k$ with 95% CI; middle column: k_1; right column: $V(\hat{N}_k)$. Each row presents a different population post-stratum. First row: Males aged 32 years with a nationality EUnew; Second row: Males aged 56 years with an Irish nationality; Third row: Females aged 28 years with an Irish nationality. All figures are rounded to nearest 10 for disclosure control purposes.

The second stopping rule considers the ratio k_1/k as possibly being convex and, if so, the stopping point will be where the bend appears to be most acute. The first case is the only one with a slightly convex curve with the bend appearing to be most acute at point $k_r \approx 520$, noting that k_1 is rounded. This stopping point is consistent with the first stopping rule for this case.

The third stopping rule relates to considering the behaviour of the variance estimate of $\hat{N}_k$ before and after k_r. The first case again is the only case where there is a case for stopping point at $k \approx 520$.

In terms of the estimates presented here, we see that trimming results in an approximate 5% reduction $(1-6460/6780)$ in the estimate for the first case, and it appears significant with regard to the 95% CI of $\hat{N}_0$. This population group relates to males of age 32 years with a declared nationality from the most recent EU countries. Ireland has experienced significant immigration in this group in recent years, and members of this group do not have a need to immediately apply for an Irish driving licence, as an existing driving licence may entitle them to drive in Ireland for a short period of time. In addition, the group may also have a relatively higher proportion of short term workers, whether on a one off or regular basis, given the ease with which it is possible to travel between EU countries. Short term workers may have no need of a driving licence but still engage with the public administration systems through paying tax. These subsets (short term workers and newly arrived immigrants) will have a relatively high probability of being trimmed. It seems therefore plausible that the set $U_A \setminus U_B$ in this population group is non-empty, which is manifested here as erroneous records in list A with respect to the joint set $U_A \cap U_B$.

The second case relates to males aged 56 years with an Irish nationality. This population group is expected to be relatively stable within the population. The third case refers to females aged 28 with an Irish nationality. The resident status can be considered more transient, due to reasons such as travel, study or work. Indeed, the set $U_B \setminus U_A$ may be non-empty for this group. One reason for this might be that the benefit of holding a driving licence may be an incentive to a small number of these persons living abroad (intending to return home shortly) to renew their driving licence on an ongoing basis. Nevertheless, the presence of such potential "erroneous enumeration" in U_B with respect to U_A would not by itself cause the rise in the TDSE.

A more plausible explanation for the different behaviour of the TDSE in the second and third case may lie with the different effects of trimming. For simplicity, suppose all the trimmed records are non-erroneous. Then, the TDSE will be higher than the untrimmed DSE, since $\hat{N}_k = n(x-k)/(m-k) > nx/m = \hat{N}_0$ as long as $m < x$. Of course, should it be the case that $\hat{N}_0 > \tilde{N} = n(x-r)/m$ to start with, we would also have $\hat{N}_k > \tilde{N}$. In other words, the trimming has already reached the stage where relatively more of the trimmed records can be found in the matched list AB among the Irish females of age 28 (before $k = 1000$) but not yet so among the Irish males of age 56 (up to $k = 1500$). Now that the TDSE is basically level to start with, there is no evidence that $U_A \setminus U_B$ is non-empty in either of these two groups based on the chosen trimming method. Notice also that the difference between the TDSE and untrimmed DSE is not significant with regard to the 95% CIs.

For an appreciation of the overall effects of trimming, we refer to the population pyramid in Figure 3.6 displaying population estimates by age and sex for 2011. The DSE $\hat{N}_0$ is given by the blue line, and the TDSE by the green line, which is based on trimming all persons with an employment record and an income less than 20k euro, denoted by $\hat{N}_T$. Notice that the DSE is only available for persons aged 17 and up, due to the nature of the *DLD*. For those aged less than 17 years old, the DSE is simply replaced by X_0 or all those identified with activity (no adjustment is made). Where $\hat{N}_0$ and $\hat{N}_T$ are the same, the graph shows the blue line overwritten by the green line. Similarly for X_0 and X_T, the grey dashed line is overwritten by the black dashed lined. This is observed in the over 65 age group where trimming is not expected to have an impact - nearly all persons

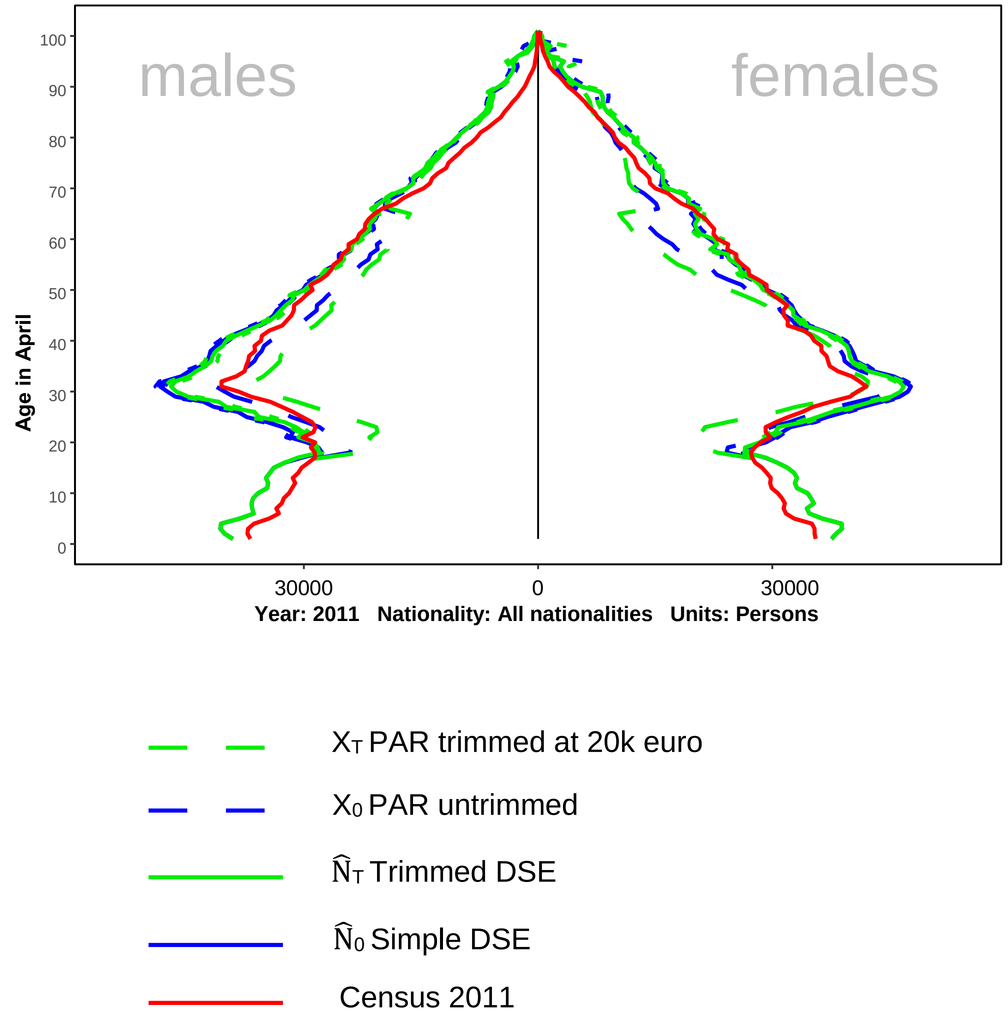

Figure 3.6: Population estimates, 2011. Impact of removing any employee that earned less than €20k. 95% confidence intervals are shown with dots around population estimates.

aged over 65 years are retired and not in paid employment. The estimates (trimmed and untrimmed) almost do not differ from each other at all for persons aged 40 - 65, despite the actual difference between X_0 and X_T, given by the blue and green dashed lines, respectively. This suggests that the set $U_A \setminus U_B$ is essentially empty in this population group. Some difference can be detected for persons below the age of 40. In particular, the TDSE of the population between age 18 and 20 is close to or slightly higher than the corresponding DSE. This is the age when many young people enter the work force and the number of trimmed records k_T is higher than the rest of population. By and large, the results suggest that the set $U_A \setminus U_B$ is nearly empty in all the relevant population groups, except for certain small groups such as in the first case presented above (32 year old males with a nationality from EU25 countries excluding EU15 countries).

While this application of TDSE was used to look for sources of overcoverage, the exercise itself raises the possibility of using TDSE to tune estimates to a particular population concept. If it is thought that $U_A > U_B$ (by definition) then TDSE can be used to trim list A such that it meets the population concept underpinning list B and U_B.

A practical application could be in estimating the population according to the usual residence concept, U_{II}, say, a duration or intention to reside for 12 months. The hypothetical SPD population, U_A, contains workers that travel to Ireland for a short period to work and pay tax but who are not part of the hypothetical population U_{II}. These workers will not consider applying for a driver licence and are also not considered part of the hypothetical DL population, U_B, which is much closer to the usually resident population concept, U_{II}. A practical application of TDSE could be to trim the SPD of all records where persons only have P35 employment and that employment is less than 20 weeks (approximately 5 months). We consider those that work for longer than this period to have resided, or intend to reside, in Ireland for 12 months. Figure 3.7 shows the impact of this trimming to be almost negligible when considering all nationalities.

3.5 Evaluation of individual data sources

3.5.1 Methodology to evaluate individual data sources

In this section the TDSE methodology will be used to evaluate the contribution of individual data sources in the SPD with respect to the compilation of the population estimates.

To evaluate a data source, the TDSE will simply have one trimming step whereby the data source is excluded from the compilation of the SPD. The evaluation is undertaken as follows:

- A first set of DSE population estimates (DSE) are compiled with the data source to be evaluated and all other data sources included in the SPD
- The SPD is then rebuilt without the data source of interest and a second set of population estimates (TDSE) compiled.
- The first set of population estimates is then compared with the second set of estimates to see if there is any difference in the estimates or their confidence intervals.

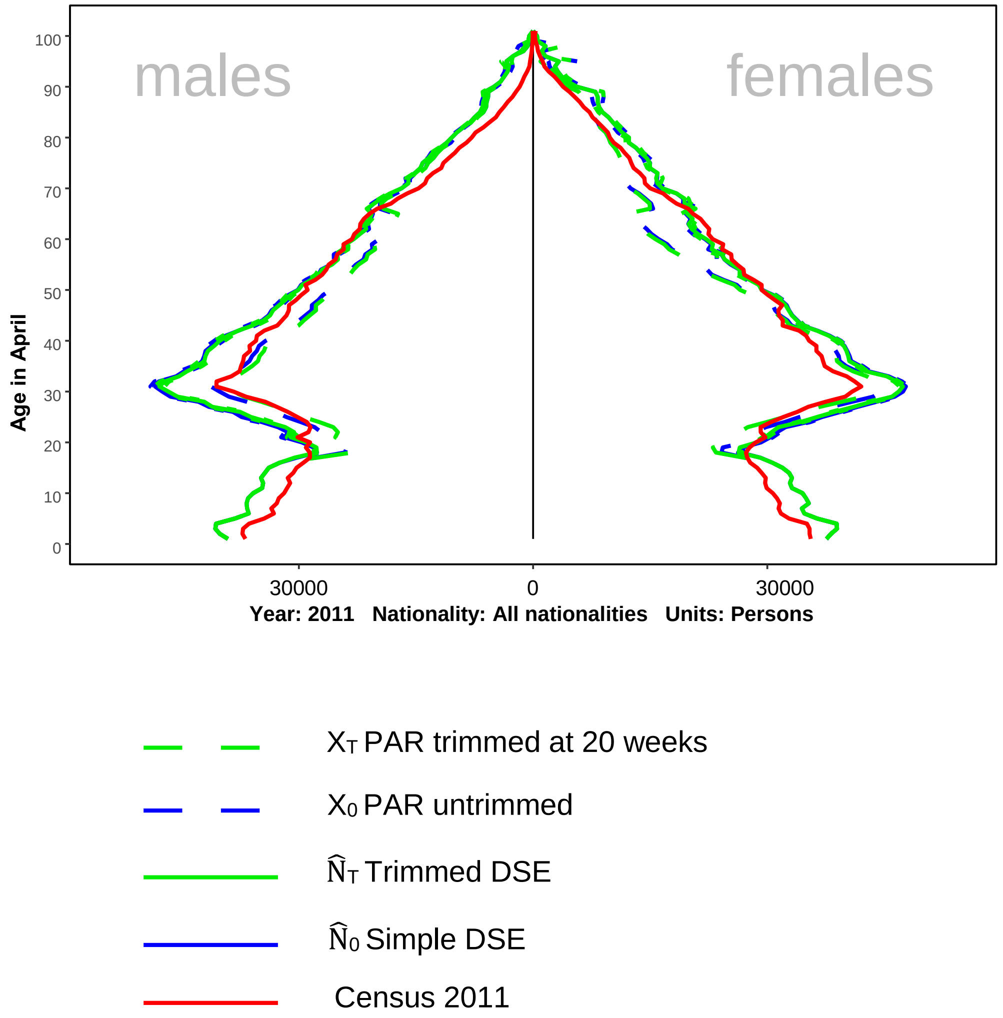

Figure 3.7: Population estimates, 2011. Impact of removing any employee with less than 20 weeks work from *PAR*. 95% confidence intervals are shown with dots around population estimates.

- If the second set of estimates is significantly lower than the first set of estimates then this would suggest the presence of erroneous records in the data source being evaluated. Consideration should be given to whether it should be included in the SPD without first looking to see how to remove the erroneous records.
- Comparing the confidence intervals of the two sets of estimates will provide an indication of what contribution the inclusion of the data source makes to the precision of the estimate.

We will use population pyramids, similar to those earlier in the chapter to evaluate the impact of a given data source on the population estimates.

Examining table 3.1 and figures 3.1 and 3.3 we will evaluate the following data sources in this section.

- Employer Employee Tax Records (P35): This data source relates to employer reports sent to the tax authorities for each employee on their payroll. The concept of an employer also includes occupational pensions. This source covers a significant part of the population.
- State Pension (SP): This is the most significant data source for those over 65 years of age. There is also a suspicion that there are erroneous records in this data source. the number of males in the older age bracket should in theory be less than the number of females when life expectancy is considered. Comparison of population estimates with the Census 2011 estimates in figure 3.3 would support this suspicion.
- Primary Care Reimbursement Service (PCRS): This data source relates to funding of health care in Ireland. The data used in this study is only available from 2013. This new data source will be evaluated to see how it adds to the system of population estimates. It is a significant data source with over 2 million persons engaging with the system every year.
- Child Benefit (CB): Child benefit payment data is a significant data source for mothers in receipt of payments. It is also critically important for estimating the population under 18 years of age.

3.5.2 Evaluation of P35 Employer Employee activity to overall system of population estimates.

Figure 3.8 presents an analysis of what happens to the population estimates when employee records are not included in the compilation of the SPD.

Exclusion of this data source has a much bigger impact on the SPD counts for males than females in terms of the reduction in the age category 20 to 65 years old. For a considerable portion of males, the SPD counts have been reduced by over 50%. This reduction translates to lower precision in the population estimates when confidence intervals are compared. Excluding the data source tends to reduce the population estimates slightly. While this difference is small it is showing up as significant in the age categories where population estimates peak, when using confidence intervals to determine significance. This hints at the possibility of some small pockets of erroneous records in this data source.

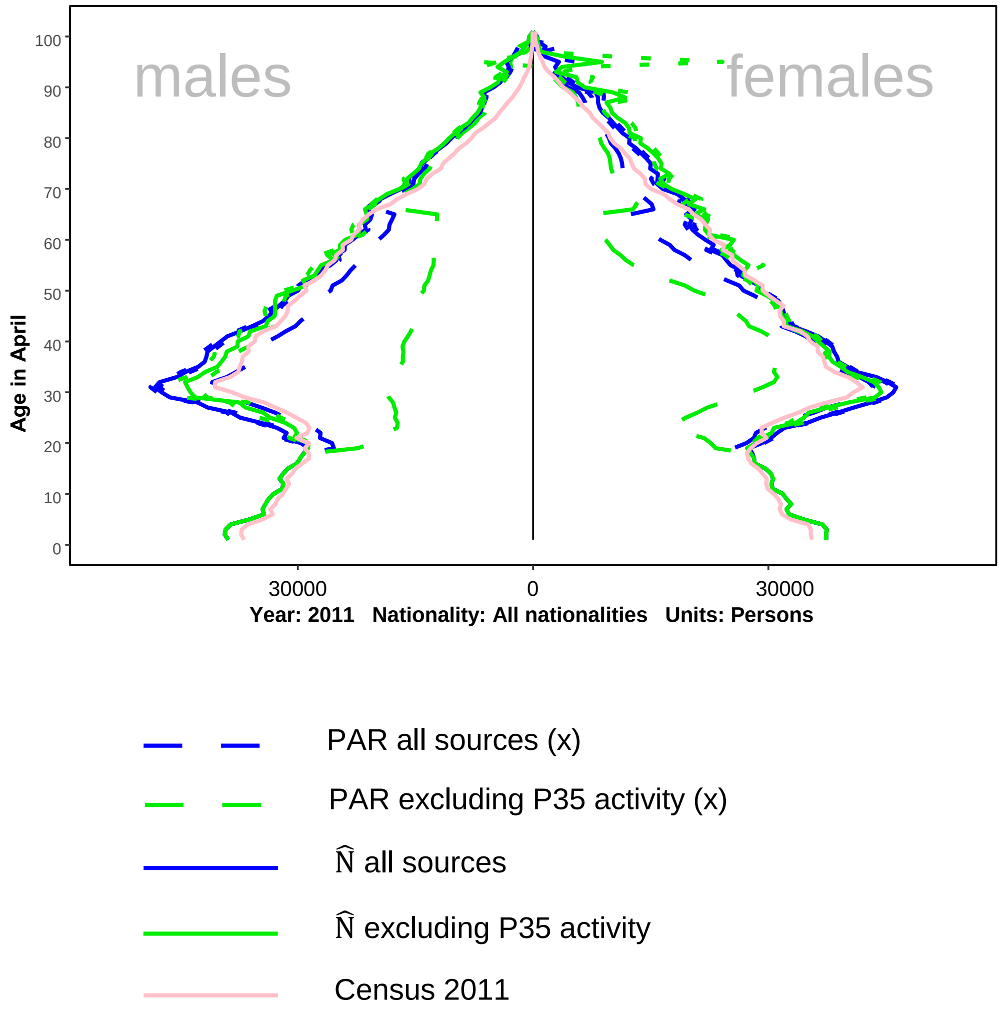

Figure 3.8: Population estimates, list B = DLD, 2011. Impact of removing P35 data source (employee records) from PAR. 95% confidence intervals are shown with dots around population estimates.

Overall, it is possible to compile population estimates without this data source. However, the contribution it makes to the SPD coverage of the population and in turn to the precision of the population estimates makes it a high value source. It should not be excluded from the SPD.

3.5.3 Evaluation of State Pension records in the overall system of population estimates

Figure 3.9 presents an analysis of what happens to the population estimates when State Pension (SP) records are not included in the compilation of the SPD.

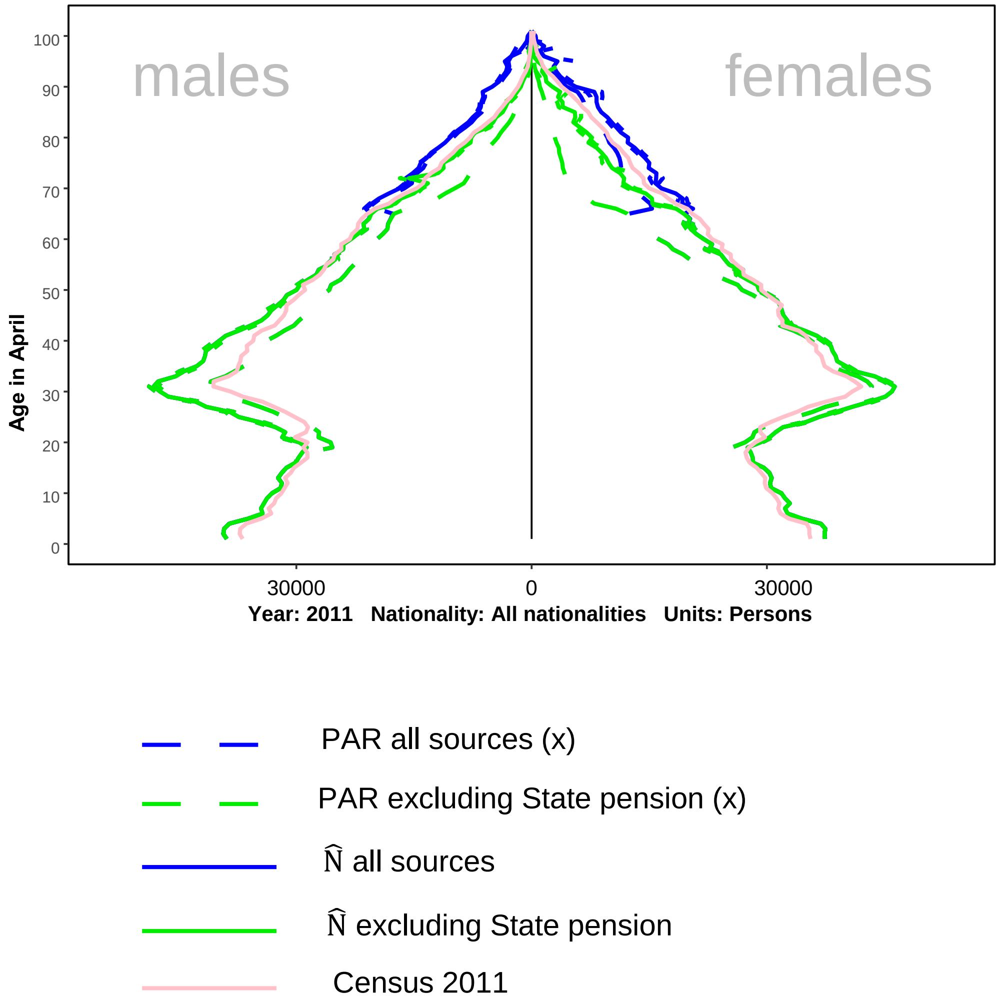

Figure 3.9: Population estimates, list B = *DLD*, 2011. Impact of removing State pension records from *PAR*. 95% confidence intervals are shown with dots around population estimates.

Trimming the State pension records from the SPD results in a sizable reduction in SPD counts in the population over 65 years of age. This reduction is significantly more pronounced on the female side of the population. Using the trimmed SPD also results in significant reductions in population estimates indicating the possible presence of erroneous records in this data source. Considering gender comparisons for the population estimates compiled from trimmed and untrimmed SPD counts, we can conclude with some conviction that there are erroneous records in this data source. It should be noted here that payment records for State pension were not available at the time of the study and a proxy indicator using different sources was created for State pension payments. It may have been the case that this proxy source did not remove persons that had passed away. At this stage there is sufficient evidence to omit this data source when compiling the SPD.

The trimmed SPD provides for significantly reduced population estimates. These population estimates look comparable with the Census population estimates for males; however, for females they are lower than the Census estimates.

We consider the significantly reduced SPD counts for females. We also give some consideration to the propensity to hold a driver licence among the elderly (and possibly among elderly women) and it may suggest that there are issues to be dealt with in this age category. We can consider what the estimates would look like under the alternative list B ($QNHS$) and evaluate the homogeneous capture assumption in this age category. This again follows the methodology presented in section 3.3, where the homogeneous capture assumption was evaluated. The presence of erroneous records in the State pension source causes a problem in validating this *homogeneous capture* assumption in the older age category, making it necessary to repeat the assumption evaluation with the trimmed SPD.

Figure 3.10 presents the same analysis as figure 3.9 except now with $QNHS$ being used as list B. There is now a significant loss in precision as $QNHS$ is much smaller in size than DLD, but $QNHS$ by definition is more likely to adhere to the homogeneous capture assumption.

When $QNHS$ is used as list B, we again see a reduction in the size of the population estimates when the State pension records are omitted, however the reduction is not as large as when DLD is used as list B. In fact, using the Census estimates as a benchmark, it looks like unbiased population estimates can be compiled, albeit at a significantly lower level of precision.

This analysis suggests that the assumption of *homogeneous capture* for the DLD data source breaks down for the older age category (those unable to pass the requisite medical check will be unable to renew a licence, a problem as older people age). Drawing from work done by Gerritse et al. (2015), whereby they demonstrate that when implied coverage (coverage of the first list, the larger list, as implied by the second list in a DSE setup) is low, deviations from independence will result in bigger deviations from the population sized estimated with fixed dependence than when implied coverage is higher, we can consider that when list counts are low there is high sensitivity to violations in the

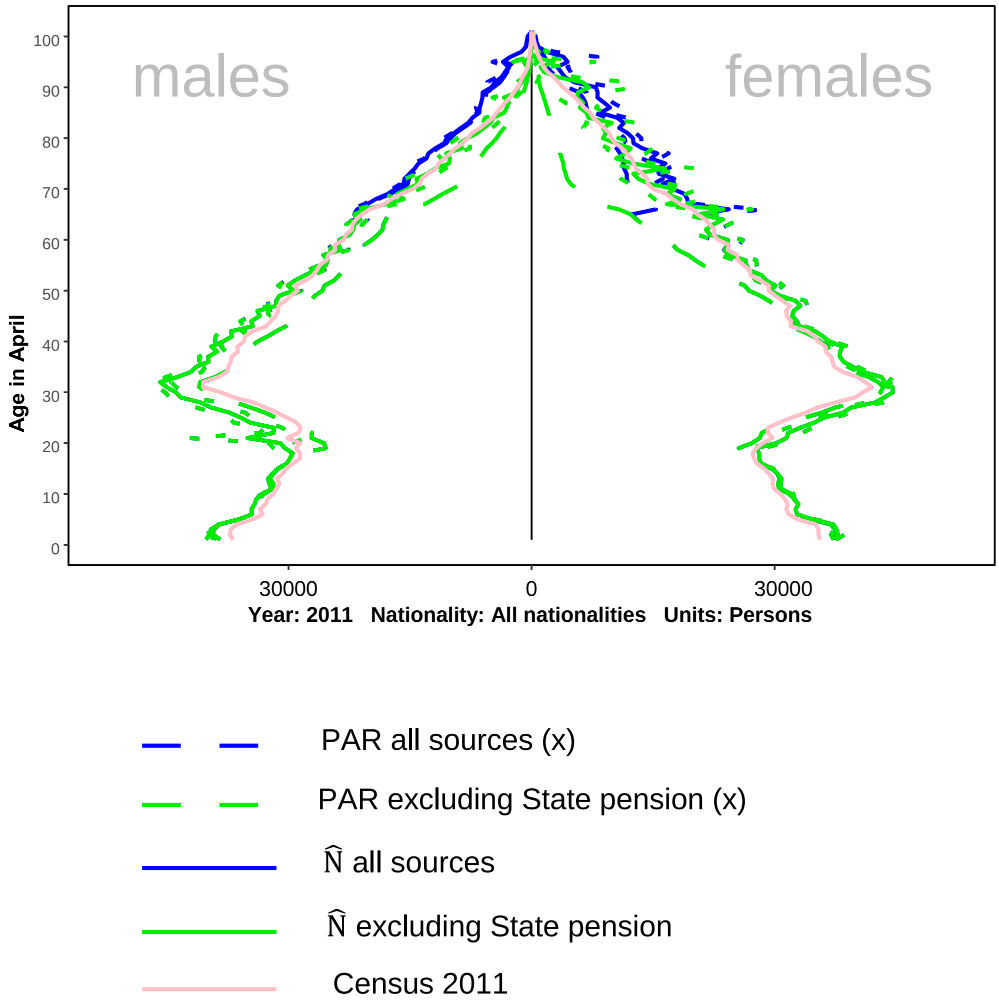

Figure 3.10: Population estimates, list B = $QNHS$, 2011. Impact of removing State pension records from PAR. 95% confidence intervals are shown with dots around population estimates.

homogeneous capture assumption for list B. Therefore, in our efforts to increase precision and accuracy in this age group, we seek to find another data source that will increase the SPD coverage rate for the population.

One such possible data source is PCRS which is available from 2013 on. Figure 3.11 presents an analysis of including PCRS data source in the SPD. Again we exclude State pensions from this analysis. Inclusion of PCRS boosts the SPD count close to the population estimate across a number of age groups beyond that of retirement age. The population estimates only differ in the level of precision and do not seem to differ otherwise. It looks like the PCRS data source is a key data source in ensuring that the SPD comes close to enumerating everybody in the State.

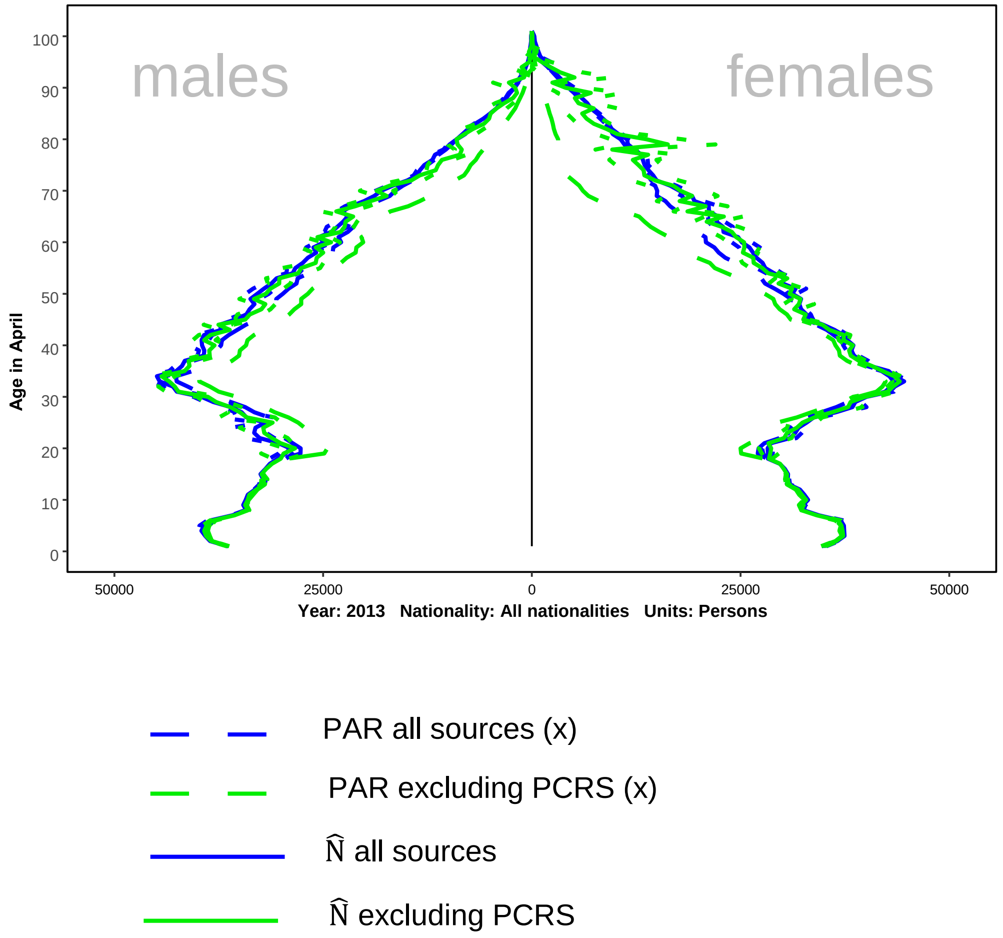

Figure 3.11: Population estimates, list B = $QNHS$, 2013. Impact of removing PCRS data source from PAR. State pensions data source is excluded from comparison. 95% confidence intervals are shown with dots around population estimates.

3.5.4 Evaluation of Child Benefit (CB) records to the overall system of population estimates

From figure 3.1 it is obvious that it is impossible to estimate the population under 12 years of age without the CB data source. These records also cover a considerable portion of the female population in the 30 to 50 years age category.

However, in 2013 a new data source (PCRS) has become available that also covers this age category. We now compare the population estimates in 2013 compiled with and without the CB data source. $QNHS$ will be used as list B to allow for adjusting of undercount in the under 18 age category. The State Pension is also excluded in the compilation of both sets of population estimates, as section 3.5.3 points to significant suspicion of erroneous records in this source. Figure 3.12 presents this analysis again using the population pyramid format.

The CB data source still has a significant impact on the SPD counts in the under 12 age category. However, over 12 years of age the Post Primary Pupils data source (PPP) compensates well when this data source is missing with almost no fall in SPD counts. The PCRS data source is the only data source that covers the under 12 year age group in the absence of the CB data source.

While the compilation of population estimates using the trimmed PAR for the under 12 age group is possible, the estimates have large confidence intervals such that they are not usable in practice. Therefore, in the absence of other data sources to compensate for the drop in SPD coverage of the population, the Child Benefit data source is critical in the compilation of population estimates.

There is only a small impact on SPD counts and no significant impact on the population estimates of the 20 to 50 year age group from removing the CB data source. This data source is retained.

3.6 Final reckoning

3.6.1 List B considerations

In section 3.3 we looked at two data sources, DLD and $QNHS$, and compared the results when used as a list B in the compilation of population estimates.

The DLD data source is a secondary data source derived from an administrative register. We make assumptions about homogeneous capture with respect to this data source. Section 3.3 then evaluated this assumption by comparing the set of population estimates with a second set of population estimates where an alternative data source, $QNHS$, is used as list B. The $QNHS$ data is a primary data collection with homogeneous capture of population units embedded in its design. This comparison showed the two sets of estimates to be, for the most part, coherent, indicating that the DLD also satisfies the homogeneous capture assumption. The set of estimates compiled using $QNHS$ will not be as precise as those compiled using DLD due to the size of the datasets.

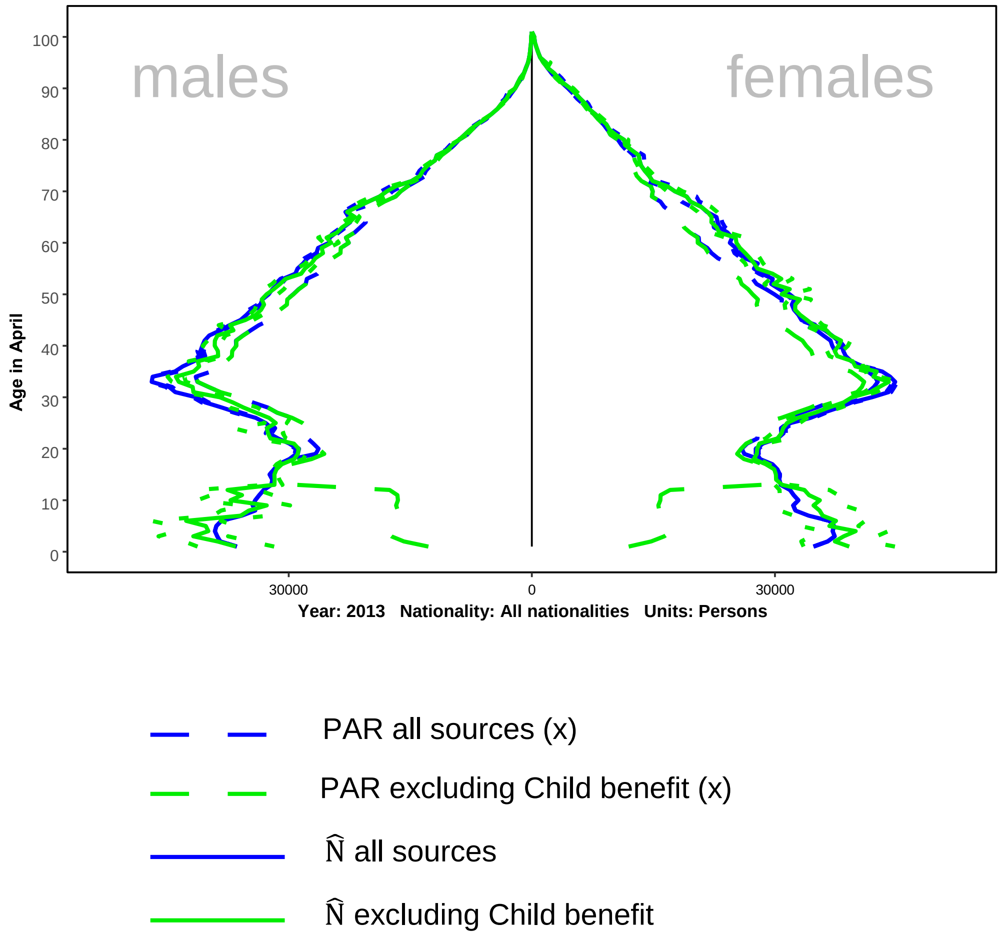

Figure 3.12: Population estimates, list B = $QNHS$, 2013, impact of removing Child benefit data source from PAR. State pensions data source is excluded in comparison. 95% confidence intervals are shown with dots around population estimates.

In a subsequent analysis presented in section 3.5.3, we suspect the State Pension data source to contain a considerable number of erroneous records. We also compile population estimates excluding State pension using two different data sources and present these in figure 3.9 which uses DLD as list B and in figure 3.10 which uses $QNHS$ as list B. A comparison of these two figures suggests that the DLD assumption of homogeneous capture does not hold for females over 70 years of age.

We make the following practical recommendations in relation to use of list B

- In general, use DLD as this will lead to more precise estimates than when $QNHS$ is used.
- For females over 70 years of age, use $QNHS$ as list B, particularly when SPD implied coverage is low, as using DLD will not fully adjust for the undercount. When the PAR implied coverage is high, the adjustment required is small and as such using DLD instead of $QNHS$ will not have a significant impact on population estimates. There is a trade off between precision and accuracy.
- If there is any reason to believe the PAR has undercoverage in the under 18 category, the $QNHS$ will need to be used as list B, as DLD will have no coverage in this part of the population. However, if no under coverage is thought to exist (SPD is considered as having complete coverage) then there is no requirement to adjust for undercoverage.

3.6.2 Data sources

In the analysis of the robustness of the proposed system of population estimates, there is evidence of erroneous records in the data source used to indicate those in receipt of State Pensions. In the absence of the actual payment data, an indicator variable is used and it is believed the erroneous records are associated with this indicator. Therefore, this data source is removed from the PAR prior to compilation of population estimates.

The precision of estimates for the over 65 years age group suffers in the years prior to the reference year 2013. In 2013 the PCRS data source became available for use. To enhance precision, the system needs to be able to include a higher quality state pension data source or PCRS data source for these earlier years. In the absence of a higher quality data source covering State Pensions, the PCRS data for 2013 is used for 2012 and 2011 for this age category with the assumption that there has been no immigration in this age category for these 2 years. Age for these years is determined based on month of birth.

PCRS is identified as a key data source for compiling population estimates from administrative data sources. It enhances the coverage of the SPD such that it can be considered close to a full enumeration of the population.

3.6.3 Generalised approach to compilation of population estimates

This work presents four innovative ideas that have been implemented to create a robust system for compiling population estimates from administrative data sources.

First, the DSE methodology is developed in a way that relaxes the traditional assumptions such that the methodology can be considered and applied in a broader context.

Second, the SPD that provides the underlying population count is created using a *signs of life (SoL)* approach. This ensures by design that the SPD suffers only from undercoverage and as such, in principle, a suitable DSE approach is all that is needed to adjust for coverage errors.

Third, instead of using a traditional Undercoverage Survey (UCS) in the field as list B in a DSE based estimate, this system of estimates uses an additional administrative data source as its list B. If it is possible to use an administrative data source as list B, then this will result in greater precision at a much lower cost. There may also be significant gains in terms of timeliness, depending on the availability of the data source.

Fourth, the DSE methodology is extended to provide the TDSE toolkit to hunt for groups of records within the SPD with proportionately more erroneous records than the SPD generally. As such, the SPD can now be trimmed to remove these problematic groups and reduce the potential for bias in the DSE based population estimate. This innovation allows for using an SPD with overcoverage provided an effective trimming strategy can be found.

The methodology and steps developed and applied in this work should not be taken and applied naively. It is better to consider them in conjunction with the underlying data sources as part of an overall strategy. Like any toolkit, the value is not in the tools and methods, but in how they are used. The strategy involves first compiling an SPD using a signs of life approach, then adjusting this SPD for undercoverage using another administrative data source as list B to obtain our population estimates. In turn then, the dataset chosen as list B is validated and the underlying SPD is interrogated with TDSE methodologies in order to adjust the system for any weaknesses and provide reassurance about the final estimates compiled from the trimmed SPD and chosen list B.

The strategy used is described in figure 3.13 where it can be seen that list B is always revalidated after trimming of the SPD and compilation of population estimates. The reason for this is that, if the SPD contains erroneous records, this in itself may have an impact in validating list B. List B is validated by a much smaller dataset compiled from a household survey with homogeneous capture inbuilt into its design. We also note the household survey design incorporates a clustering feature which will have an impact when considering variance estimation.

3.6.4 Results

In summary, the following decisions were made in finalising the set of population estimates.

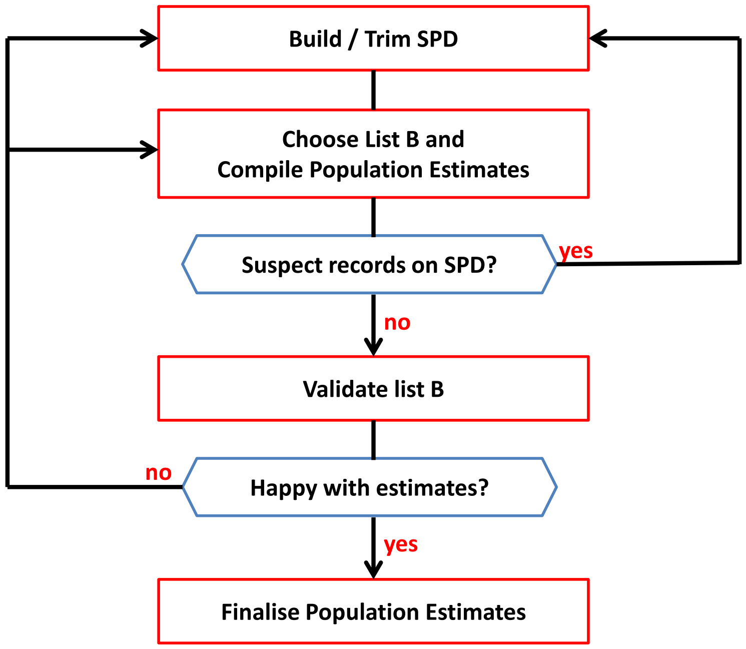

Figure 3.13: High level process map for compilation of population estimates.

- The 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 in the upper age categories.

- The PCRS records (health related records) for 2013 were aged appropriately and included as a data source in the SPD for 2012 and 2011 to counter the poor coverage in these years for the older age categories. 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. There is sufficient coverage of the retirement age category 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 the P35 data source before it is included in the SPD. This, in theory, has the effect of tuning the estimate, such that the underlying population concept equates to that of an *annual resident population* (Lanzieri, 2013), and excludes temporary or migrant workers who may come and work for a period of 20 weeks or less. This ensures the underlying population concept is better aligned to the commonly used concept of usual residence (12 months residence, intended or actual).

The population estimates, along with precision estimates, for years 2011 to 2016 are presented in table 3.2. Coverage of the SPD for 2016 drops slightly as the income tax returns for the self employed and the higher education enrolment (HEA) data were unavailable for 2016 at time of compilation (see table 3.1, for details on data source availability).

A comparison of the population estimates and Census usual resident counts by gender for 2011 and 2016 are provided in table 3.3 and this comparison, broken down by age is presented using population pyramids in figures 3.14 and 3.15.

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 UR count occurs for young adult males between the ages of 20 and 40 years old.

There are 4 possible explanations for the difference between the population estimates and the Census UR count.

The first explanation relates to the underlying population concept or definition. The Census UR count is a count of 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. It is reasonable to equate the resident concept of the population estimate with the usual resident concept of the Census UR count. The primary difference between the two concepts relates to the Census UR count relating to a specific night while the population estimate can relate to any night in the calendar year. This would imply that to equate the Census UR 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 1.1 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. Therefore, we can say conceptual differences between the two measures will account for 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 no erroneous records on the DLD. The DLD has also been validated in section 3.3 as a list B data source. While this validation focussed on the homogeneous capture assumption, the validation should in theory not be successful if DLD had 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 3.3) generally validated this assumption there were some indications of a small violation of the homogeneous capture assumption for females in the older age category. This violation may be seen in figure 3.9 where, 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 $QNHS$ in figure 3.10, the population estimate moves back above the Census estimate, albeit with a larger confidence interval.

The fourth and final 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. Northern Ireland recorded an undercount of 4.7% in the 2001 Census (Abbott, 2009) and a similar undercount in 2011 (NISRA, 2015). The undercount was significantly greater for young adults. Given the information and analysis in this chapter and findings with undercoverage surveys in Northern Ireland with respect to the Census, it is reasonable to consider the possibility of undercoverage in the Irish Census. We consider the possibility of undercoverage in the Irish Census further in chapter 4.

3.7 Concluding remarks

The concept of the Census is changing. Given the number of countries involved in modernisation through the use of administrative data sources, the Census is now becoming defined by its outputs rather than by its process (UNECE, 2015). Historically, the population concept has been derived in a manner that best fits a traditional Census with the population being counted or enumerated where they live on a specific night. With a number of countries successfully implementing a register based Census, the population concept has been broadened to allow for the population count to be based on registered persons, with geography being decided on the legally recorded address for individuals

Year	2011	2012	2013	2014	2015	2016
List A (SPD) - Number of final records used						
	4,397,770	4,424,370	4,533,430	4,541,630	4,611,800	4,473,900
SPD Coverage of population (percent)						
	91	92	93	92	92	89
List B (*DLD*)						
	422,680	507,030	468,870	378,100	466,610	539,200
Match between list A and list B (*DLD*)						
	376,950	452,730	425,130	341,230	422,070	462,330
Population estimate						
	4,811,020	4,828,990	4,896,230	4,925,380	4,992,260	5,038,640
CV of Population estimate (percent)						
	0.05	0.05	0.04	0.05	0.04	0.05

Table 3.2: Population estimates compiled from administrative data sources, 2011 to 2016. Published on http://www.cso.ie/shorturl/480 as research outputs, December 2018

	Population Estimate	Census (usual resident)	Difference	Difference (%)
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

Table 3.3: Comparison of PECADO Population Estimates with Census usual resident counts by gender, 2011 and 2016. Published on http://www.cso.ie/shorturl/480 as research outputs, December 2018

(UNECE, 2014). Lanzieri (2013) considers population definitions in the context of the increase in variation of definitions as countries modernise and proposes an *annual resident population* concept. Lanzieri argues that this new concept is easier to implement in the context of the broader use of administrative data sources. With the possibility of a requirement to meet different population concepts it is important that any system has the ability to be able to tune the estimates to match these concepts. In the work here, we have tuned the estimate to meet the concept of *an annual resident population.*

The Irish case has a significant advantage in that there exists a high quality system of official identifiers to remove the problem of linkage error. The choice of *DLD* as list B ensures high quality identification numbers due to the nature of authentication with respect to renewing or applying for a driver licence. List A is also compiled from administrative data sources where a transaction or significant service has been provided

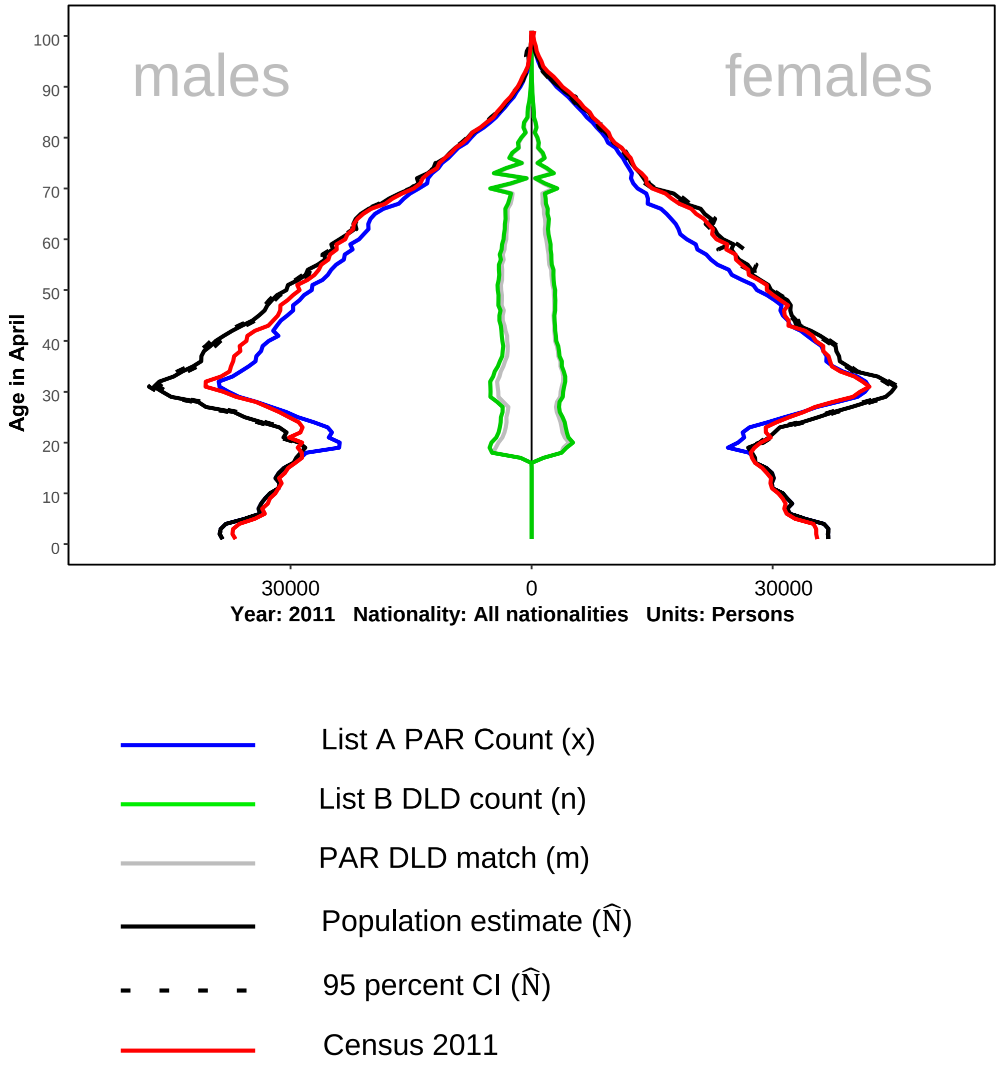

Figure 3.14: Comparison of final population estimates and Census usual resident counts by age and sex, 2011

and this in itself also provides for confidence in the quality of the identification numbers in the underlying data sources. In the absence of personal identification numbers, the system would then have to deal with the highly likely situation of linkage error leading to false positives and false negatives in terms of identifying the match between list A and list B. False positives (negatives) will lead to a negative (positive) bias in the population estimate. In this scenario, list B is better fulfilled by a field survey to identify coverage issues in a manner similar to that conducted in the 2011 Spanish Census (Argüeso and Vega, 2014) where both undercoverage and overcoverage are addressed.

In the presence of linkage errors we can consider the true match m as $m = m_L - e + u$

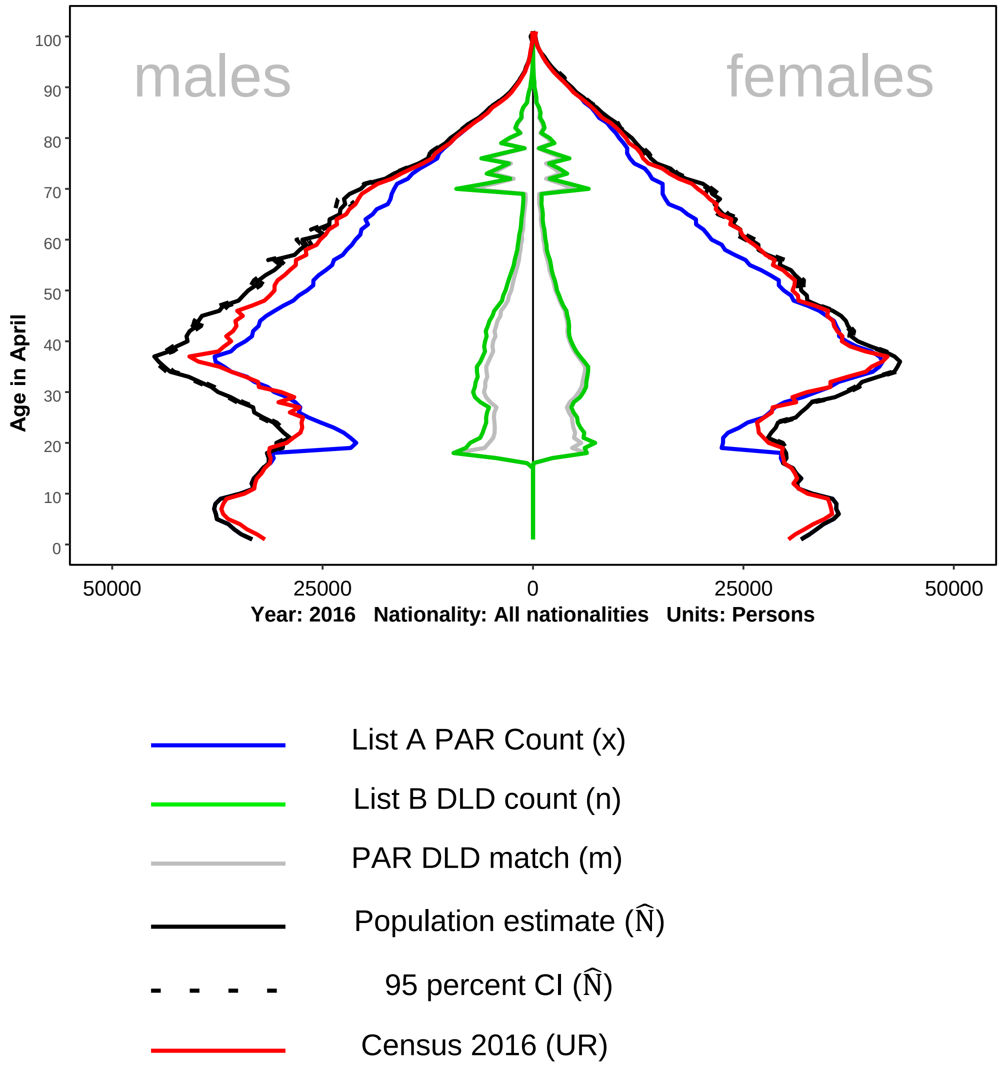

Figure 3.15: Comparison of final population estimates and Census usual resident counts by age and sex, 2016

where m_L refers to the linked records or the observed match, e represents the number of false links (false positives) made and u represents the number of missed links (false negatives). In the chosen trimming strategy, we remove a data source and compare population estimates before and after trimming to see if there is any significant difference. If there is no difference and assuming perfect linkage, we have no evidence of erroneous records in the trimmed data source from this trimming step. If there is a difference and we consider no erroneous records are present in list A this would indicate a possible change in the nature of the linkage error before and after trimming and as such provide evidence of linkage error with either the part of the list removed by trimming or the trimmed list (or both). While we have not developed the DSE methodologies in the presence of linkage error, the stability and robustness of the population estimates to further trimming provides added comfort that with respect to the absence of linkage error. Zhang and Dunne (2018) also notes the significant challenge of working with DSE applications in the presence of linkage error.

To the authors knowledge, this is the first example of a system of national population estimates compiled solely from administrative data sources without using a Central Population Register. The SoL approach reduces the number of problems from four to one (undercoverage) to be addressed by the system. This allows for a simple application of DSE methodology that is relatively easy to explain to users. There is significantly higher value in using activity or SoL based data than using registration information. This suggests that NSIs should be focussing more on negotiating access to activity or SoL data to compile population estimates.

Chapter 4

Census Undercoverage Survey (UCS) using administrative data sources

4.1 Background

CSO, Ireland to date has not conducted a Census UCS. CSO has relied on the effectiveness of its operations and the diligence of all staff involved (in particular field staff) to conduct the Census without a requirement to undertake a UCS to adjust for undercount. Generally, undercoverage is a feature of the traditional Census. The undercoverage rate in the 2011 Northern Ireland Census is reported as 5% after additional active persons on administrative data sources have been included (NISRA, 2015). Statistics New Zealand reported the net Census undercount as 2.4% in the 2013 Census (Statistics New Zealand, 2014a). 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).

CSO has committed to undertaking a UCS as part of the Census in 2021. If conducted in the traditional way, the additional cost of a UCS is thought to be in the region of €0.6m. The largest part of this cost is associated with a second field operation.

In this chapter, we explore the possibility of a UCS using administrative data sources. If feasible, this possibility would negate the need to conduct a second field operation and the associated costs. We use the 2016 Census of Population dataset to evaluate this proposal. Undercoverage, if present, may also explain some of the difference between the population estimates compiled from administrative data sources and Census counts presented in chapter 3.

4.2 Methodology

The methodology is based on the DSE methods as developed in section 2.2.2.

In the typical Census operation with an associated UCS, the Census is considered as list A (size x), the followup UCS is conducted to compile list B (size n), and then list A and list B are linked to identify the match (size m). The assumptions followed are those described in Wolter (1986). A simple DSE equation such as that in equation 2.4 (page 43) is then used to obtain $\hat{N}$, an estimate of the size of the population U. The undercount in the Census is simply obtained by differencing. Post stratification is incorporated to estimate undercoverage in specific groups and to relax the assumptions with respect to capture and independence such that they only need apply within strata.

In deriving the DSE estimator in section 2.2.2, a different set of assumptions are used such that the multinomial assumption can be relaxed and replaced with an assumption of *homogeneous capture*, or the assumption of a constant probability for a unit of the population being captured in a list only has to apply to one list - list B. Chao et al. (2008) also explain how Wolter's assumptions can be relaxed, such that *equal catchability* only has to apply to one list, and make particular reference to potential applications for Census coverage surveys.

This now allows the undercoverage survey for the Census to be set up as follows:

Denote the Census list as list B of size n where each unit in the population has an equal chance (probability close to 1) of being captured in the Census. Let list A of size x be drawn from an administrative source such that any unit in list A

1. meets the criteria of the population concept that the Census is estimating and
2. can be clearly identified as enumerated by the Census or not, that is there is no *linkage error*.

The statistical unit can, in theory, be either dwellings or persons. In this application, the statistical unit is persons. An estimate of the undercount on the Census is now obtained using the DSE methodology. Post-stratification by covariates is used to deal with any heterogeneity in capture rates across groups and any required disaggregations where the covariate information is available and consistently recorded for each unit on lists A and B.

This approach is different to the *Reverse Record Check* component in the Canadian Census (Statistics Canada, 2015). In the *Reverse Record Check* approach, a list B sample is selected from a frame compiled using the previous Census dataset updated with administrative data sources (births, deaths, migration); the target population of the frame is the same as that for the Census. The sample is then matched with the current

Census dataset in a DSE exercise. A first phase of matching is done via record linkage (automated and computer assisted) to accurately classify as many as many records as possible without having to contact persons directly. Those records from the sample that cannot be easily classified are contacted in a second phase follow up field exercise to determine whether these residual records were counted in the Census or not and if not, whether they should have been included in the frame for sample selection. The Canadian approach is underpinned by the traditional assumption of independence between two lists. However, there may be an opportunity to use the PECADO approach to DSE to rethink the Canadian approach such that the the requirement for the field survey can be eliminated. The following summary steps outline how the requirement for a field survey in the Canadian setup can be eliminated.

1. Consider the current Census dataset as list B meeting the homogeneous capture assumption.
2. Consider the previous Census dataset updated with administrative data sources as list A
3. Match list A to list B using existing automated and computer assisted methods and classify each record in list A as *Matched*, *No match* and *Unsure* as to whether the record has a match in list B.
4. Trim list A of all records classified as *Unsure* to obtain a trimmed list A.
5. Compute population estimates using the trimmed list A and list B in the DSE setup as outlined.

This approach could be evaluated/validated by using Census 2011 as a test bed. This approach may even be superior to the existing approach where sampling error is introduced into the subsequent classification of unsure records. This approach should also be easier to understand and explain to users from a methodological perspective as there would no longer be a requirement to conduct a survey to deal with records that have been classified as unsure.

4.3 Data

In April 2016, CSO enumerated 4.74 million people as usually resident and present on Census night in the Republic of Ireland. We will denote this estimate as the Census UR count. This Census did not incorporate a UCS, however, some imputation of individuals was undertaken based on information collected in the field where it was not possible to directly enumerate persons known to be resident in a dwelling. This typically occurred where an enumerator was unable to obtain a direct response from a household but indirectly was able to collect some information from a neighbour. This resulted in the final Census dataset of 4.74 million records, a record for each person usually resident and present on Census night.

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 create enhanced statistical products through linking the Census data with administrative data. The first output published using this new linking capability focuses on household income (Visit https://www.cso.ie/en/releasesandpublications/ep/p-gpii/geographicalprofilesofincomeinireland2016/incomeinireland/, last accessed on 28th June 2020).

Relevant safeguards were deployed in creating the linkage capability between the Census dataset and administrative data sources. These safeguards involved separating Census identity information (name, date of birth, address, gender) from Census attributes and then matching this identity data set against a similar identity dataset created from PPSN registration information. Once the links were created between the identity datasets match keys were created and deployed in place of the identity information on the original dataset. The CSOPPSN, or Protected Identifier Key (PIK) for the PPSN, was also deployed on the Census dataset thus allowing linking with administrative data sources while protecting the identity of underlying persons. The safeguards also included separation of tasks among different teams so no team had access to identifiable information as well as sensitive attribute information such as Income, Welfare payments etc. The team that undertook the actual linking of identifiable components (or the identity set) of the datasets used a combination of automated matching and computer assisted matching. 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 the 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 where a PPSN has been identified, 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. Furthermore, when using this dataset in a DSE setup we will assume *homogeneous capture* for each person in the population on Census night. This means that any potential undercoverage and any inability to link persons captured in the Census with administrative data sources is assumed to be missing at random (MAR) within post-stratification blocks.

List A is based on a dataset (DSP payments), which contains a record for each person receiving a Social Welfare payment in April 2016. This list also contains the same identifier key, CSOPPSN, as on the *trimmed* Census dataset. The assumption is that each person receiving a welfare payment in this month was usually resident and present on Census night. This assumption is reasonable 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.

Both list A and list B have high quality identification keys based on official identification Numbers. Any linkage error, if it exists, is assumed to be negligible.

4.4 Analysis and Results

After blocking by single year of age, nationality grouping and gender the results are plotted in figure 4.1 for single year of age and gender. The administrative list, plotted with a green continuous line, comprises of social transfers consisting of unemployment related benefits and pension payments. It has very poor coverage in the below 20 age group and as such we will discount the DSE based model estimates in this category. The *trimmed* Census count is plotted with a red continuous line while the match between the two datasets is plotted using a green dashed line. The estimated usual resident population or UCS population estimate, under the DSE model, is then plotted in blue with confidence intervals plotted using dashed lines. An adjustment has been made to the UCS population estimate prior to plotting as follows; where the list A count is below 50 for the point plotted or the DSE estimate itself is below the observed Census UR count, the DSE estimate is replaced with the Census UR count.

The UCS population estimate (blue) is then plotted in figure 4.2 along with the Census UR count (green). We will refer to this count as a preliminary UCS estimate, as later, we make an adjustment in the under 20 years age category to get a final estimate. The final PECADO population estimates (red) presented in section 3.6.4 are also included for comparison purposes. We note again the differences in the 3 underlying population concepts. The Census UR count relates to being usually resident and present with the usually resident component relating to being resident or having an intention to be resident for 12 months or more on Census night. The UCS estimate relates to a population concept of being usually resident in the month of April without the requirement of also being present. Qualifying for a welfare payment requires being able to prove that you are usually resident in the state without necessarily having a 12 month limit. The PECADO population estimate is best aligned to an annual resident population type concept where a person has had a significant period of residence (say, greater than 6 months) and has been resident at any point in the calendar year.

Assuming no difference due to population concepts and that assumptions hold, we will interpret the difference between the Census UR count and the UCS estimate as undercoverage. Figure 4.2 shows some undercoverage in the Census UR count across all age groups over 20 and for both genders. The undercoverage is far greater for the male population of working age with the undercoverage declining as the age moves towards the retirement age at 65. The UCS estimate is set to the Census UR count where the list B coverage is less than 50 persons in any gender by single year of age group. We note here the MAR assumption (*homogeneous capture*) for list B covers 733,800 persons, or nearly 15% of the final UCS population estimate which is a much larger proportion of the population to be estimated for than in the typical Census setting with undercount.

In practice, there will be a very small number of people not present but usually resident in April; for example, those persons out of the country on holidays or for work purposes.

So in summary, while some undercoverage exists across all age categories over 20 years of age, this undercoverage looks to be greater for males of working age. This observation is in line with experiences in other countries that conduct a traditional Census and an associated UCS.

In comparing the UCS estimate with the PECADO population estimates in figure 4.2, the estimates look to be well aligned except for the 20 to 40 age categories for both genders where the PECADO estimates are relatively higher. Migration flows may reasonably explain the differences between the PECADO estimate and the UCS estimate. To conceptually derive the UCS estimate (usually resident in April) from the PECADO estimate (usually resident at any given time in the calendar year), emigration in the period January to March and immigration in the period May to December must be subtracted from the PECADO estimate. From table 4.1, the difference between the UCS and PECADO estimate for the age group 20 and over is 37,000 persons. In considering annual migration flows (immigration plus emigration) of approximate 140,000 from table 1.1 (page 2) and the slight discrepancy in the over 60 years age group for both sexes, the difference in the 20 to 40 age group is reasonably expected.

There is also a similar but less obvious difference in the under 10 year age category for both genders. In considering the under 10 age category, the PECADO population estimates use a simple count of administratively active persons. It is not possible to evaluate undercoverage for this age group using driver licence data in the PECADO project while the UCS is a straight count of the Census UR population as again there is no data available in the list B to adjust these counts. So, assuming no erroneous records for this age group in the PECADO project, the conclusion is that there is undercoverage in this age category in the Census UR count as migration flows will be negligible.

From figure 4.2, there is also a slight but consistent discrepancy between the UCS estimate and the PECADO estimate. Recalling the results from the PECADO project in section 3.6.4 (page 90), we note the existence of a probable violation of the homogeneous capture assumption among females aged over 70 in the Driving Licence Dataset (list B) resulting in a negative bias in the PECADO population estimate. We also noted that this bias is small if the coverage of list B is high and this bias may now explain the

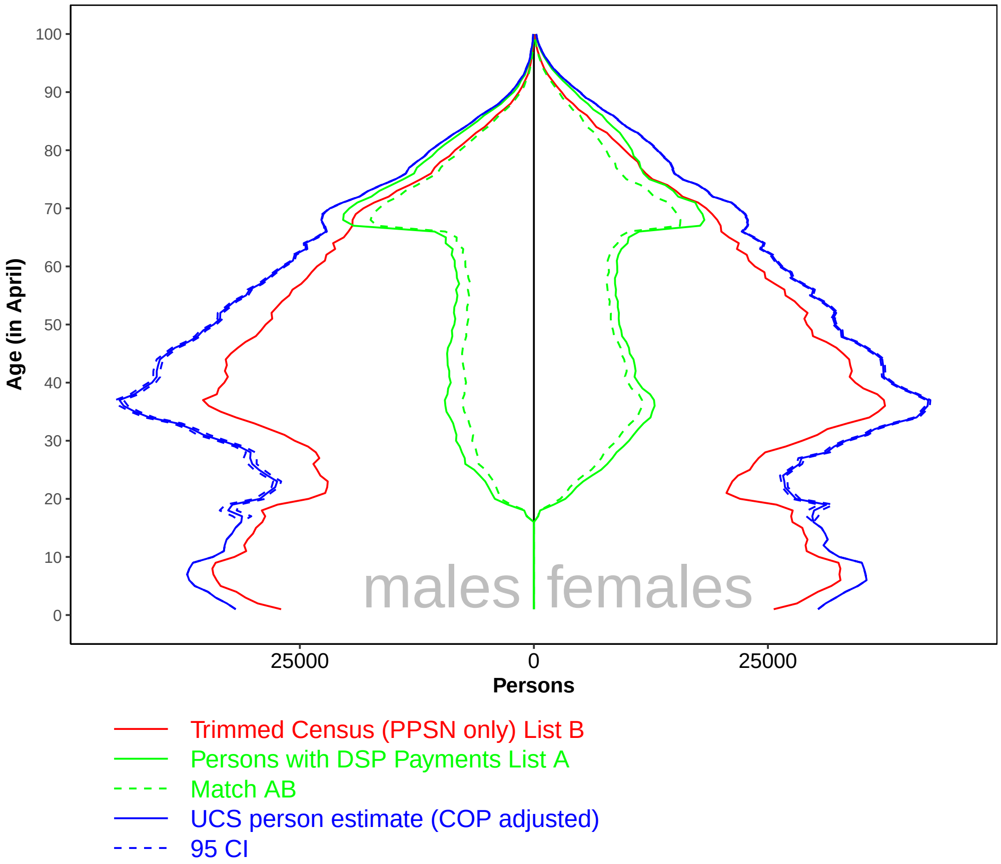

Figure 4.1: UCS usual resident population estimates compiled using DSE, 2016. A *trimmed* Census dataset satisfying list B conditions is used as list B while an administrative list based on Social transfers is used as list A.

difference between PECADO estimate and the UCS estimate for persons in the over 60s age category.

To obtain an overall estimate of undercoverage in the Census UR count, we derive a final UCS estimate for those usually resident in April by adding the PECADO estimate for the under 20 year age group to the UCS estimate for the over 20 years age group, to obtain an estimate of 5,001,700 persons. Comparing this figure with the Census UR count of 4,739,600 gives an estimate of undercoverage of approximately 262,100 or 5.2%. Adjusting the preliminary UCS estimate with the PECADO estimate for the under 20 age group is justified as there is strong evidence these persons were resident in the State in the reference year and a negligible likelihood of migration.

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 5% being derived through a coverage assessment and adjustment process. In this context, an estimate of undercoverage at 5% is relatively good. However, it does point to the fact that CSO, Ireland can

	Age group Under 20	20 and over	All ages
Census (Usually Resident)	1,306,700	3,432,900	4,739,600
Trimmed Census	1,193,000	3,074,900	4,267,900
DSP payments	15,400	1,397,300	1,412,700
Match	13,300	1,160,000	1,173,300
UCS Population Estimate (Preliminary)		3,668,300	
PECADO Population Estimate	1,333,400	3,705,300	5,038,600
UCS Population Estimate (Final)	1,333,400	3,668,300	5,001,700
Implied Census Undercoverage (%)	2.0	6.4	5.2

Table 4.1: Comparison of Census Usual Resident counts, Preliminary Census Usual Resident estimates adjusted for undercoverage and PECADO population estimates by age group, 2016. The preliminary UCS estimate only considers those aged 20 and over as the Welfare Payments data source is restricted to adults. The final UCS population estimate uses the PECADO estimate for under 20 year olds.

no longer rely on accepting at face value that the current implementation of the traditional Census model can eliminate undercoverage. Some form of UCS will be required to provide the necessary reassurance that the Census does not suffer from undercoverage. Also, CSO, Ireland needs to consider how to develop and deploy methods that will allow Census counts to be adjusted for undercoverage if it exists.

4.5 Conclusions and Discussion

The main conclusion to be drawn from this chapter is that CSO has made a smart decision to include a coverage survey in 2021. The CSO should no longer assume that its traditional Census model, relying solely on population counts, does not underestimate the Usually Resident population at Census night without providing reassurance to this effect. This chapter provides some evidence that Census 2016 contained an undercount. Given the assumptions underpinning the UCS and the size of the adjustment between the *trimmed* Census and UCS estimate in table 4.1, it is prudent to treat this evidence as circumstantial until assumptions can be further evaluated.

When comparing population concepts, it should be noted that there is a small difference between the UCS population concept and the Census (Usually Resident) population concept. The hypothetical UCS population is slightly bigger than the Census (Usually resident) population. The difference is generally accounted for by persons who emigrate in the month of April prior to Census night plus persons who immigrate in the month of April subsequent to Census night. There are also a very small number of people who are resident but not present on Census night (those away for work or holiday reasons for example). The difference between the Census (Usually Resident) population concept and the PECADO population concept is much larger. The PECADO population concept

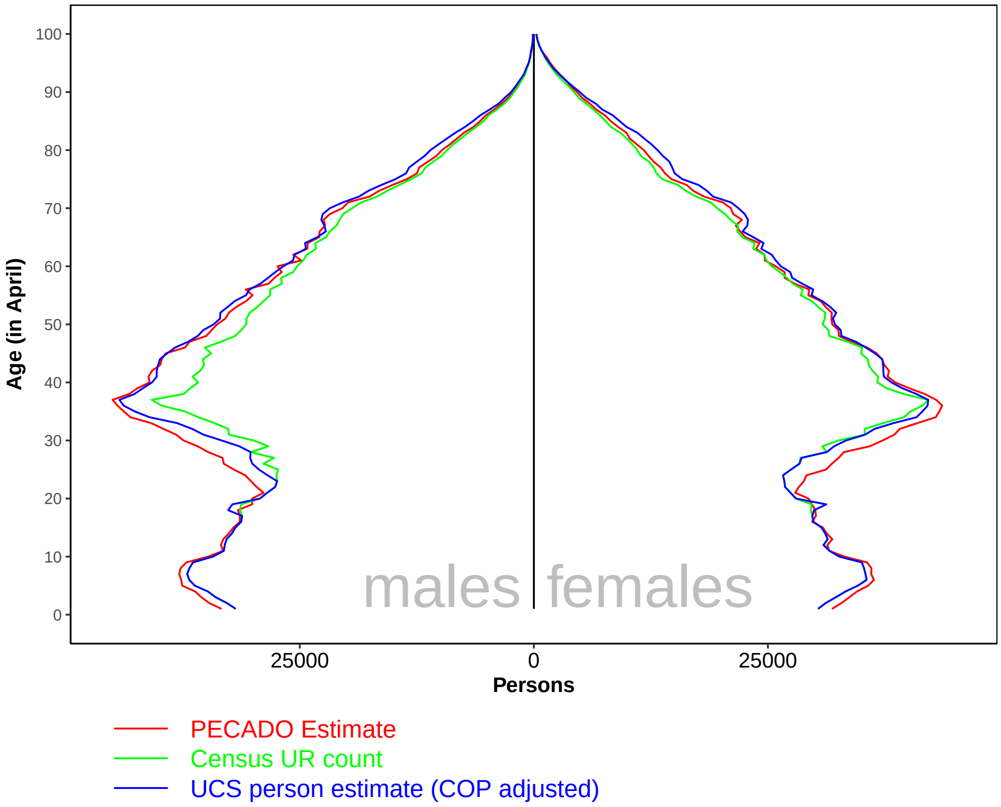

Figure 4.2: Comparison of Census Usual Resident counts, Preliminary Census Usual Resident estimates adjusted for undercoverage and PECADO population estimates by gender and single year of age, 2016

includes anyone resident at any time in the reference year while the Census (Usually Resident) concept refers to Census night only; therefore to derive the Census (Usually Resident) population from the PECADO population estimates, emigration from the start of the year to Census night plus immigration from Census night to the end of the year must be subtracted.

The second conclusion is, with only one list in the DSE model required to satisfy the *homogeneous capture* assumption, it may be possible to replace the traditional costly survey deployed in the field with a suitable administrative data list. Provided the following assumptions hold: no linkage error, every person on the administrative list also belongs to the population and the Census list satisfies the list B conditions (i.e. homogeneous capture). Furthermore, taking on board the findings in section 2.4, the homogeneous capture assumption can also be relaxed under certain conditions.

This chapter only explores estimates of undercoverage at a State level. In practice, estimates of undercoverage broken down by geography will also be required. The project also briefly explored this possibility. The approach taken used dwelling and person within dwelling as the match key to link list A and list B. Given that enumerators visit every dwelling in the 2016 Census, we can assume the geospatial information on the Census dataset is of high quality. In this exercise, list A was restricted to only those records which had high quality geospatial information for the dwelling, that would enable an unambiguous link (or not) between the two lists. We found significant inflation of our population estimates due to geographical misplacement or address mislocation between list A and list B. This can occur for a number of reasons, the obvious example being an elderly person who may be living with family or in a nursing home but who maintains an administrative presence at the place they considers their usual residence. It may be possible to extend the DSE or TDSE methods to adjust for address mislocation. The idea behind the solution is that the State level estimates would act as a benchmark for calibration purposes. We refer to the extension of methods in the PECADO toolkit in this manner as a calibrated DSE (CDSE) or calibrated TDSE (CTDSE).

The further development of DSE methods in this manner will have considerable value. The CDSE/CTDSE methods would also have application to the PECADO population estimates as similar issues are expected to arise in this context also.

Chapter 5

Estimation of population flows when Statistical Population Datasets (SPDs) are incomplete

5.1 Introduction

5.1.1 Background and Context

For those countries that do not have a Central Population Register on which demographic statistics can be derived, the production of reliable demographic statistics on population counts and migration flows can prove challenging. This is particularly true for those countries that have high and variable migration flows.

The typical approach to population estimates in these scenarios is an application of the demographic component or cohort component method. This method is also extensively used for population projections. The approach 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. This approach is typically applied across the different 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 and New Zealand undertake a Census at 5 yearly intervals. The Census provides a benchmark to recalibrate the population estimates at regular intervals. Table 1.1 (page 2) in chapter 1 provides an overview of the estimated population and the change components for Ireland over the years 2011 to 2016.

In Ireland, the principal source of information for the estimation of the gross annual migration flows has traditionally been the Quarterly National Household Survey (QNHS), which also provides the basis for the classification of the flows by sex, age group, origin/destination and nationality. The QNHS targets 25,000 households (approximately 2% of households in the state) each quarter. The QNHS was replaced by a new Labour Force Survey (LFS) in Quarter 2, 2017. The LFS and QNHS have similar survey design characteristics. The migration estimates are also compiled with reference to movements in other migration indicators such as the number of Personal Public Service Numbers (PPSN, Irish PIN) allocated to non-Irish nationals and the number of visas issued to Irish nationals with respect to a number of destinations including Australia, US and Canada. In addition, data on National Insurance Numbers (UK PIN) issued to Irish nationals in the UK is considered. More information on how CSO, Ireland currently estimates migration can be found at https://www.cso.ie/en/methods/surveybackgroundnotes/populationandmigrationestimates/ (accessed on 24th September 2018).

The relative size of migration flows (just under 3%, table 1.1) and the QNHS sample size ($< 2\%$) present considerable challenges in estimating migration flows. As migration flows currently contribute to the compilation of population estimates, any new developments that can enhance the quality of these estimates will have significant value in their own right.

In chapter 3, we proposed and evaluated a new system of population estimates that does not depend on first estimating the components of change or having population estimates for the preceding years. This system of population estimates is based first on compiling a Statistical Population Dataset (SPD), called the Person Activity Register (PAR), which contains one record for each person summarising their administrative activity (pay tax, enrolment in education, receive welfare, etc.) in a given year. The PAR uses a *Signs of Life* (SoL) approach summarising evidence that someone has engaged with the state and, as such, is resident in the state in a given year. By definition the PAR, if used for estimating the population in a given year, is designed to have undercoverage and no overcoverage. Therefore, the PAR is considered an incomplete list of all persons living in the state in a given year. In deriving population counts, DSE (Dual System Estimation) methods are deployed to adjust for undercoverage. The second list used is another administrative data source (not included in the PAR) that satisfies the assumptions underpinning DSE models. This second list is created from the Irish driver licence database and contains a list of persons who either applied for or renewed their licence in a given year. This driver licence renewals (including first time applications) list is called the Driver Licence Dataset (DLD). We will refer to the system of population estimates from chapter 3 as the PAR/DLD method in this chapter. The PAR/DLD method does not include estimation of gross population flows.

In reviewing the work of others, only one solution considering the extension of DSE methods that has an application in estimating population flows when lists are incomplete was discovered (Chao et al., 2008). Chao et al. (2008) have previously explored how to estimate the *shared population* using capture-recapture methods. We will refer to their approach as the Chao method. As we will see later, their extension of the Petersen-Lincoln estimator cannot be directly applied to the PAR/DLD system of population estimates due to a shortcoming in the relationship between the data sources. However, we consider it here, as it may have applications in similar situations where the underlying data sources do not suffer the shortcomings in the PAR/DLD method. The Chao method requires an overlap between all four data sources. The Chao method is illustrated in figure 5.1 and is discussed further in section 5.1.2.

In a further consideration of the situation where the 4 lists overlap, a Quadruple System Estimator (QSE) could also in theory be a solution. The use of multiple system estimators relies on each cell in the cross classification of the lists having a non zero value to estimate the non observed part of the population in the lists, where the lists are not considered to provide complete coverage of the population. In this scenario, the 4 data sources would be categorised as a *2 year population based approach*. More information on these methods can be found in the textbook by Bishop et al. (1975).

5.1.2 Chao Method to estimate Stayers

Chao et al. (2008) consider an extension of the Petersen-Lincoln estimator or DSE methods to estimate the size of a shared population, which can be equated to the population of Stayers or units that belong to the population at two time points. Figure 5.1 illustrates the Chao method if the 4 lists overlap. The Chao method can be presented as a simple DSE method where the first list relates to those persons common to PAR in two consecutive years ($PAR_1 \cap PAR_2$), shaded green and yellow in figure 5.1, and the second list relates to those persons common to the list of driver licence renewals in two consecutive years ($DLD_1 \cap DLD_2$), shaded green and pink in figure 5.1. The denominator of the DSE is the intersection shaded green in figure 5.1 and can be expressed as the set $DLD_1 \cap DLD_2 \cap PAR_1 \cap PAR_2$. Chao et al. (2008) also show that the equal catchability or homogeneous capture assumption does not need to hold for the first list - in our application, $PAR_1 \cap PAR_2$.

Unfortunately, the Chao method is unlikely to be useful in estimating the population of Stayers using the PAR/DLD setup. From our previous description of Driver Licence rules (see section 1.3.2.2), we know that persons will not typically renew their driver licence in consecutive years; therefore, the second list, $DLD_1 \cap DLD_2$, will contain

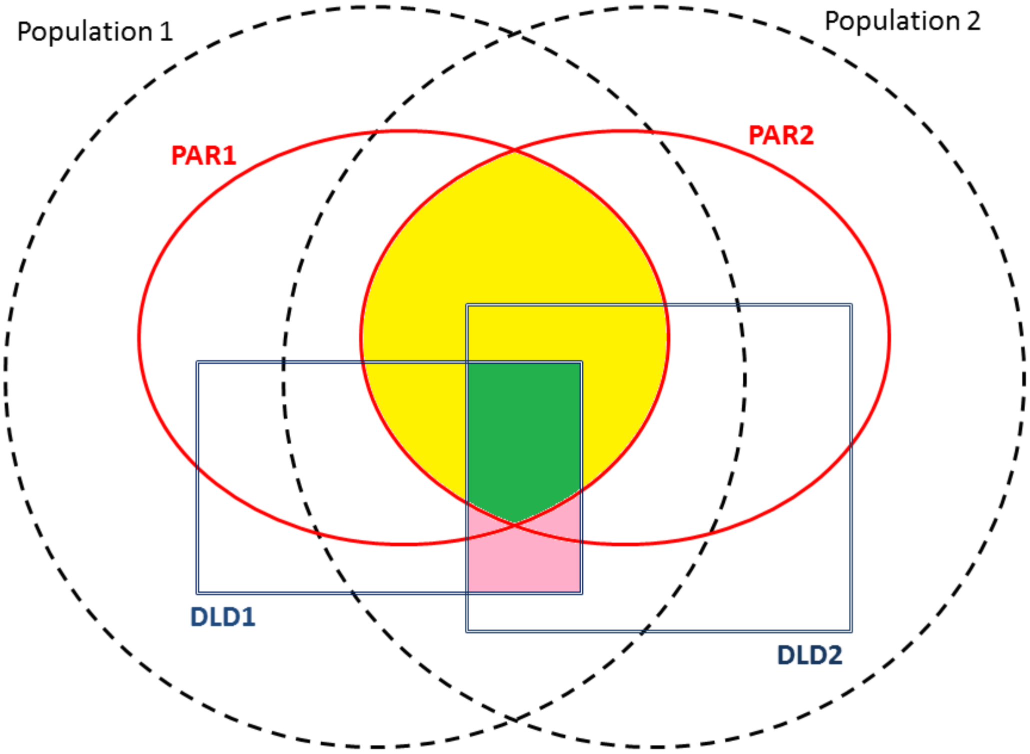

Figure 5.1: Estimating Stayers using Chao method. Depends on all 4 lists overlapping.

low counts, leading to a vastly reduced match and consequently to instability in the estimator[1]. However, if the DLD component of the PAR/DLD method is replaced by another administrative data source that has an appropriately sized overlap in consecutive years, the Chao method (and similarly the QSE approach) may be applicable.

Given the potential low numbers in those their driver renewing licence in consecutive years, we develop the underlying methodology under the assumption that $DLD_1 \cap DLD_2$ is null.

5.1.3 Outline of Chapter

This chapter proposes an extension to the PAR/DLD method that will allow for estimation of population flows, that, is how to estimate population flows when population

[1] An analysis of 2015 and 2016 Driver Licence activity data across age groups for 2015 and 2016 indicates between 3% and 11% of those that renewed their licence in 2016 also renewed or applied for their licence in 2015. For those aged over 85 this number could rise significantly higher - 30% of the small number of 95 year olds issued their licene in 2016 were also issued a licence in 2015, presumably a reflection of the rules around medical certificates of competence for older drivers

registers or SPDs are incomplete. If the PAR/DLD method provides for reliable population estimates in consecutive years, then reliable estimates for net population change can be obtained by simply differencing estimates in two consecutive years. If reliable estimates of gross population flows (inflows and outflows) can also be obtained, then reliable estimates of migration flows are also obtained, as official figures for births and deaths are readily accessible and compiled from administrative data sources. Immigration and emigration estimates are obtained by subtracting the number of births and deaths from estimates of population inflows and outflows respectively.

The remainder of the chapter is laid out as follows:

- Section 5.2 proposes a preferred method to estimate Gross population flows using an extension to the PAR/DLD method to estimate population sizes
- Section 5.3 considers alternative approaches to the PAR/DLD extension proposed in section 5.2
- Section 5.4 explores some of these different methods using data simulations and compares them to the proposed method.
- Section 5.5 presents the results of our proposed methods when applied to real world data in the Irish context
- Section 5.6 finishes the chapter with some concluding remarks.

Table 5.1 (page 138) provides an overview of the proposed and alternative methods discussed when estimating population flows.

5.2 Proposed Migration Methodology

5.2.1 Methodology Overview

We consider, as a starting point, the relationships in the population over two consecutive years. In order to simplify notation, we consider consecutive years, and use numbers 1 and 2 to denote those consecutive years. U_1 and U_2 are used to denote the two populations. The Stayers in the population (those in the population for two consecutive years) can be defined as $U_S = U_1 \cap U_2$. We can also consider a *2 year population* or any person that belongs to the population in either year or both years as $U_{12} = U_1 \cup U_2$. Persons belonging to the Outflows sub-population U_O can now be defined in relation to both U_{12} and U_S as $U_O = U_{12} \setminus U_2 = U_1 \setminus U_S$. Persons belonging to the Inflows sub-population can also be defined in relation to U_{12} and U_S as $U_I = U_{12} \setminus U_1 = U_2 \setminus U_S$. We also use N_1, N_2, N_{12}, N_S, N_O and N_I to denote the population size for each of U_1, U_2, U_{12}, U_S, U_O and U_I. Figure 5.2 visualises these relationships.

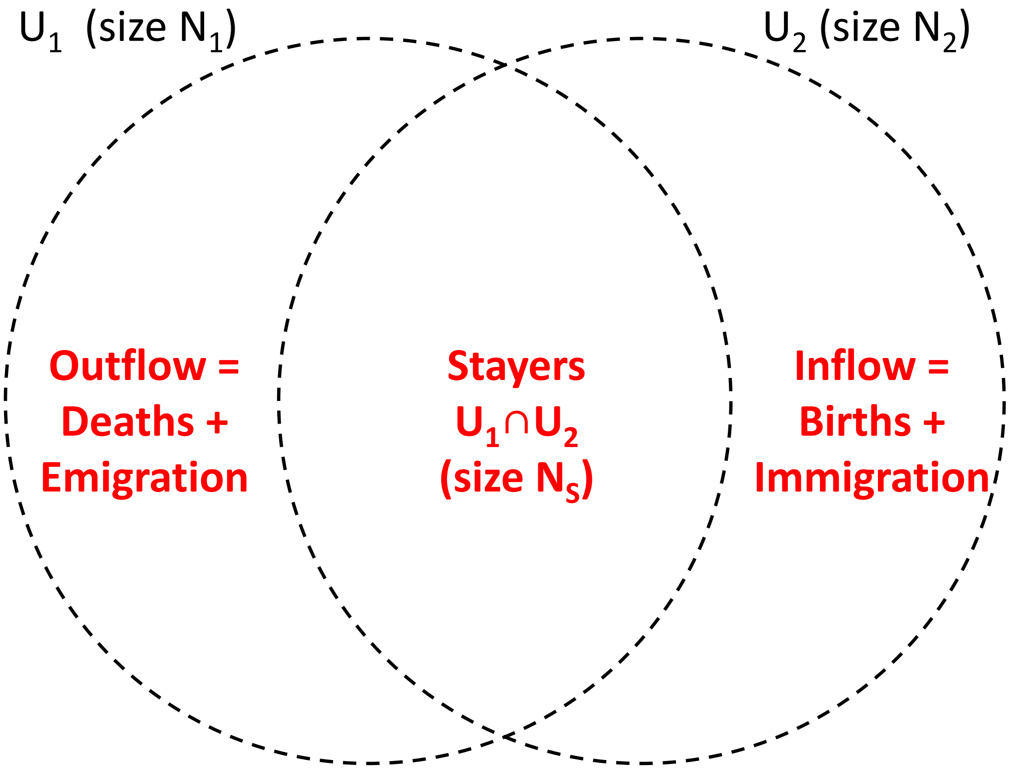
U_1 (size N_1)
U_2 (size N_2)
Outflow = Deaths + Emigration
Stayers $U_1 \cap U_2$ (size N_S)
Inflow = Births + Immigration

Figure 5.2: Illustration of relationship between Outflows, Stayers and Inflows within a Population presented for two consecutive years. U_1 denotes the population for year 1 and U_2 denotes the population for year 2.

The proposed approach to extending the PAR/DLD method to allow for the estimation of population flows is based on first estimating N_S and then obtaining estimates for N_O and N_I by differencing with N_1 and N_2 respectively. The PAR/DLD method already provides estimates of N_1 and N_2.

The data sources used as part of the PAR/DLD method to estimate the population size are also the only data sources used in extending the PAR/DLD method to estimate the population flows. We consider DLD_1 and DLD_2 as *representative* samples of U_1 and U_2 respectively and assume they satisfy the *homogeneous capture* assumption for their respective populations. Heterogeneity in capture rates is addressed using appropriate post stratification strategies. We now take figure 5.2 and overlay the 4 data sources (PAR_1, PAR_2, DLD_1 and DLD_2) to illustrate how the data sources and their intersections relate to the populations in figure 5.3. Figure 5.3 is critical to understanding the proposed method (and alternative methods). Given the assumption of no erroneous linkage between the data sources, we can directly observe and enumerate the 4 data sources and any intersections between them with confidence.

Given the data sources, it has not been possible to identify a method to estimate the Stayers directly. However, it is possible to obtain an estimate of the Stayers in the population indirectly by first estimating the Stayers in the PAR for either year and grossing or scaling this estimate to the population. Given that the PAR will have high coverage of the population it is anticipated that any errors due to the violation of additional assumptions will be small and go to zero as the coverage of the PAR goes to the population. This last point is a key consideration in proposing this approach over other identified approaches.

In summary, the proposed method to estimating gross population flows is as follows:

1. Estimate the Stayers in the population over 2 consecutive years, years 1 and 2. This is done indirectly by first estimating the Stayers in the underlying PAR and then scaling up to get an estimate of Stayers in the underlying population. year 1 or year 2 can be chosen for this approach.

2. Gross population flows are now obtained by differencing the Stayers in the population N_S with the given population estimates for years 1 and 2, N_1 and N_2.

5.2.2 Estimating Stayers in the PAR (SPD) - a DSE based approach

A broader consideration of the Chao method leads to the consideration of a DSE approach to estimate the number of Stayers contained in SPD for year 1, where the two lists to be used are $PAR_1 \cap PAR_2$ (list A) and $PAR_1 \cap DLD_2$ (list B). Figure 5.3 (page 116) illustrates how the estimator is compiled, with $PAR_1 \cap PAR_2$ shaded green and pink, and $PAR_1 \cap DLD_2$ shaded yellow and green. We will refer to this estimator as the DSE based estimator for Stayers. Note we can also apply the methodology in reverse and create a DSE based estimator for the number of Stayers contained in the SPD for year 2, where the two lists to be used are $PAR_1 \cap PAR_2$ and $PAR_2 \cap DLD_1$.

If we apply this proposal to our situation we obtain a DSE based estimate of the number of Stayers contained in PAR_1 ($\hat{K}_1$) as

$$\hat{K}_1 = \frac{x_{1:2} n_{1:2}}{m_{1:2}} \tag{5.1}$$

where

$\hat{K}_1$ is a DSE based estimator of the total number of Stayers from the population in both years 1 and 2 that are contained in PAR_1 compiled using DLD_2,

$x_{1:2}$ is the number of persons in the PAR for both years 1 and 2,

$n_{1:2}$ is the number of persons in the PAR for year 1 and in the DLD for year 2,

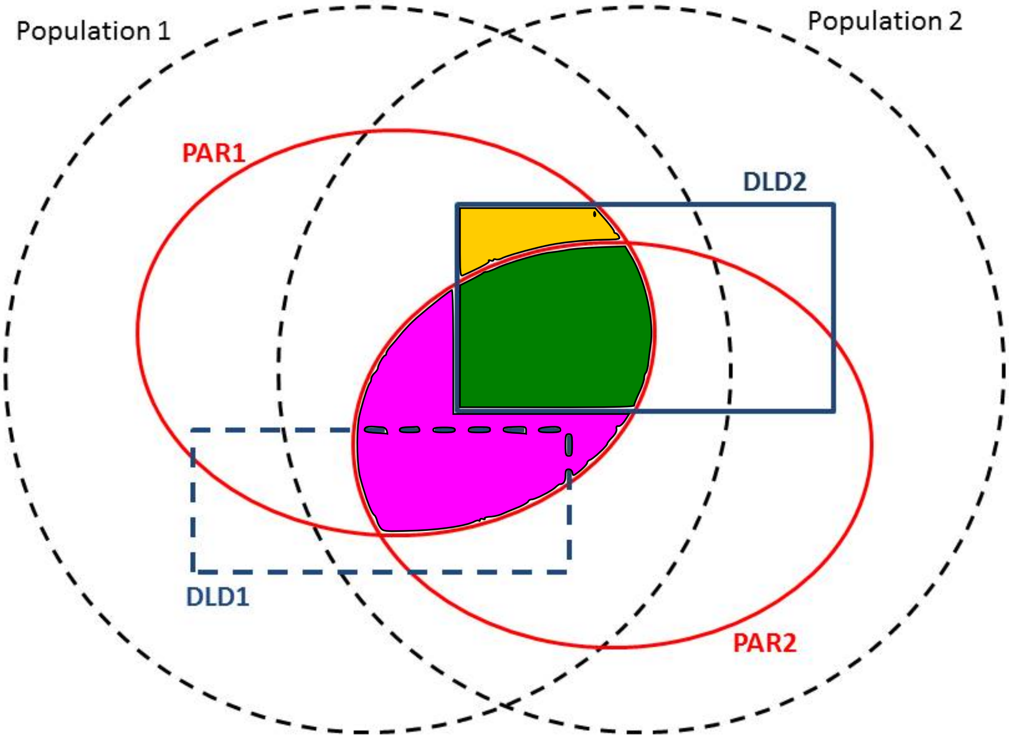

Figure 5.3: DSE estimate for Stayers contained in PAR_1. Using the two lists $PAR_1 \cap PAR_2$ (green and pink shaded area) and $PAR_1 \cap DLD_2$ (yellow and green shaded area) can provide an estimate for the number of Stayers in PAR_1, that is the number of persons from PAR_1 that would appear in Population 2.

$m_{1:2}$ is the number of persons in the PAR for both years 1 and 2 and in the DLD for year 2.

An estimate of the variance of the estimator of $\hat{K}$ can be obtained as follows

$$\hat{V}[\hat{K}_1] = \frac{x_{1:2}n_{1:2}(x_{1:2} - m_{1:2})(n_{1:2} - m_{1:2})}{m_{1:2}^3} \tag{5.2}$$

The precision of the estimator $\hat{K}_1$ is driven by the size of the match $m_{1:2}$ and the size of this match with respect to the lists $x_{1:2}$ and $n_{1:2}$. In particular, if $m_{1:2}$ is small the estimator will be unstable.

In a similar manner, the DSE based estimator for the number of Stayers contained in PAR_2 using DLD_1 ($\hat{K}_2$) can be written as.

$$\hat{K}_2 = \frac{x_{2:1}n_{2:1}}{m_{2:1}} \tag{5.3}$$

where

$\hat{K}_2$ is a DSE based estimator of the total number of Stayers in the population between years 1 and 2 that are contained in PAR_2,

$x_{2:1}$ is the number of persons in the PAR for both years 1 and 2 and is the same as $x_{1:2}$,

$n_{2:1}$ is the number of persons in the PAR for year 2 (PAR_2) and in the DLD for year 1 (DLD_1),

$m_{2:1}$ is the number of persons in the PAR for both years 1 and 2 and in the DLD for year 1.

Again, we can obtain an estimate of the variance of $\hat{K}_{2(DSE)}$ as

$$\hat{V}[\hat{K}_2] = \frac{x_{2:1}n_{2:1}(x_{2:1} - m_{2:1})(n_{2:1} - m_{2:1})}{m_{2:1}^3}$$

In general, application of the DSE in this situation will conform to the assumptions in section 2.2.2, where the DSE methodology is developed. Taking each of the assumptions in turn:

Homogeneous capture: Each unit (person) in the population of interest (say, $PAR_1 \cap U_2$ for the forward based DSE estimator), has equal probability of being caught or included in the corresponding DLD list (DLD_2 for the forward based DSE estimator). The assumption underpinning the population estimates for year 2 is that each unit belonging to the population U_2 has an equal probability of being included in DLD_2. As $PAR_1 \cap U_2$ is a subset of U_2 by definition, the assumption also holds for the DSE application here. The same logic also applies to the backwards based estimator.

No linkage error: There is no linkage error between the two lists. As for the case in producing population estimates, linking is deterministic and relies on high quality official identification numbers being attached to each unit, and as such the assumption can equally be argued to hold in this instance.

No erroneous records: Both lists, PAR and DLD have been defined and built in a manner such that they do not contain overcount. The lists have also been built based on high quality identification numbers attached to records with significant evidence that a unit belongs to the respective population.

5.2.3 Estimating Stayers in the population

The method now proposes grossing up the number of Stayers in the PAR to the number of Stayers in the population for year 1. The grossing factor is simply $\hat{N}_1/x_1$ where $\hat{N}_1$ is the estimated population size in year 1 and x_1 is the number of persons observed in

PAR_1. This method requires an additional assumption that a person's propensity to migrate has no relationship to the likelihood of being included in the PAR. If the PAR coverage is high, any violation or weakness in this assumption may have only a minor effect, as the grossing factor will be close to 1.

The estimate can be written as (choosing either year 1 or year 2)

$$\hat{N}_S = \frac{\hat{K}_i \hat{N}_i}{x_i} \tag{5.4}$$

where $\hat{K}_i$ is the estimate of Stayers in PAR_i

$\hat{N}_i$ is the estimate of the population in year i

x_i is the size of the PAR in year i

Noting a comparison of the exact variance of the product of random variables with the usual approximation (Goodman, 1960), we use the usual approximation

$$V[AB] = E[A]^2 V[B] + E[B]^2 V[A] + 2E[A]E[B]Cov[AB] \tag{5.5}$$

to develop a variance estimator for $\hat{N}_S$ in equation 5.4.

$$\hat{V}[\hat{N}_S] \approx \frac{1}{x_i^2}\left(E[\hat{N}_i]^2 V[\hat{K}_i] + E[\hat{K}_i]^2 V[\hat{N}_i] + 2E[\hat{N}_i]E[\hat{K}_i]Cov[\hat{K}_i, \hat{N}_i]\right) \tag{5.6}$$

We have already considered estimators for $V[\hat{K}_i]$ (equation 5.2) and $V[\hat{N}_i]$ (equation 2.5) so that only leaves consideration of the term $Cov[\hat{K}_i, \hat{N}_i]$ to obtain a variance estimator for $\hat{N}_S$, equation 5.6.

$$Cov[\hat{K}_1, \hat{N}_1] = Cov[\frac{x_{1:2} n_{1:2}}{m_{1:2}}, \frac{n_1 x_1}{m_1}] \tag{5.7}$$

We can identify three covariance terms that are necessary to be able to determine $Cov[\hat{K}_1, \hat{N}_1]$. The three covariance terms are $Cov[n_{1:2}, m_{1:2}]$, $Cov[n_{1:2}, m_1]$ and $Cov[m_{1:2}, m_1]$ where

$n_{1:2}$ is the size of the set $PAR_1 \cap DLD_2$

$m_{1:2}$ is the size of the set $PAR_1 \cap PAR_2 \cap DLD_2$

and m_1 is the size of the set $PAR_1 \cap DLD_1$

In considering $Cov[n_{1:2}, m_{1:2}]$, we partition $n_{1:2}$ into two parts as follows

$$n_{1:2} = \sum_{i \in (U_2 \cap PAR_1) \setminus PAR_2} \delta_i + \sum_{i \in (PAR_1 \cap PAR_2)} \delta_i$$

corresponding to the yellow and green shaded areas respectively in figure 5.3 and considering $m_{1:2}$, the shaded green area in figure 5.3, as

$$m_{1:2} = \sum_{i \in (PAR_1 \cap PAR_2)} \delta_i$$

where $\delta_i = 1$ if $i \in DLD_2$ and $\delta_i = 0$ otherwise. We also note the homogeneous capture assumption with respect to DLD_2 for U_2 and can write $P(\delta_i \in DLD_2) = m_2/x_2$. We also note the independent capture assumption for each person. We see that the second part of $n_{1:2}$ is the same as $m_{1:2}$ is the same and, as such, we obtain an expression for $Cov[n_{1:2}, m_{1:2}] = Cov[m_{1:2}, m_{1:2}] = V[m_{1:2}]$ and using the binomial distribution to write

$$\widehat{Cov}[n_{1:2}, m_{1:2}] = x_{1:2} \left(\frac{m2}{x_2}\right)\left(1 - \frac{m2}{x_2}\right)$$

While we can determine an expression for $Cov[n_{1:2}, m_{1:2}]$, attempts at finding suitable expressions for $Cov[n_{1:2}, m_1]$ and $Cov[m_{1:2}, m_1]$ have proved challenging and unwieldy.

However, we argue the case that the covariance term $Cov[\hat{K}_1, \hat{N}_1]$ should be small and will be negative, and as such, setting it to zero to estimate $V[\hat{N}_S]$ is a reasonable and conservative approach. The source of variation in the estimation of K_1 comes from DLD_2 (size n_2) while the source of variation in the estimation of N_1 comes from DLD_1 (size n_1). Drivers only renew their driver licences every 10 years and typically do not renew their licence in consecutive years. Based on these typical rules and to demonstrate the argument, we consider the population as being assigned to 10 cohorts with equal probability to represent the year in which they are due to renew or apply for their licence. In this scenario, we can use the multinomial distribution with parameters $(N, p_1 ... p_{10})$ where N is a notional population size to speculate the the covariances are small as follows

$$\begin{aligned} E[n_k] =& Np_k = N(0.1) \\ V[n_k] =& Np_k(1 - p_k) = N(0.1)(0.9) \\ Cov[n_j, n_k] =& -Np_j p_k = -N(0.1)(0.1) \end{aligned}$$

The covariance term here is negative. The argument therefore is that if we consider the covariance term to be zero then the following is a useful and conservative estimator of the variance of N_S. The quantity will be larger than that presented for the variance of the estimator in equation 5.6.

$$V[\hat{N}_S] \approx \frac{1}{x_i^2}\left(E[\hat{N}_i]^2 V[\hat{K}_i] + E[\hat{K}_i]^2 V[\hat{N}_i]\right) \tag{5.8}$$

We note that using year 2 instead of year 1, in a similar manner as outlined above, will provide a second estimate of Stayers in the population. In theory, the best combination would be a weighted estimate based on the relative variances of the two estimates to be combined. However, in order to keep the compilation simple, we propose a final estimate as a simple average of the two estimates.

5.3 Consideration of Alternative Approaches

5.3.1 Estimating Stayers in the SPD based on the Hypergeometric Distribution

5.3.1.1 Method Development

In this section, we consider an alternative approach where instead of using DSE methods to estimate the number of Stayers in the PAR, we develop and consider an approach using the Hypergeometric distribution.

The hypergeometric function is used to calculate the probability of k successes in n independent draws, without replacement, from a population of size N that contains exactly K successes. The probability of $X = k$ successes can generally be written as

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{5.9}$$

Adapting to the problem of estimating the number of Stayers in the PAR, we consider the probability of finding m_{12} Stayers in n_2 independent draws (DLD_2), without replacement, from population 2 of size N_2 that contains exactly K_1 persons that are also in PAR_1. So K_1 is the number of Stayers in PAR_1. m_{12} refers to $PAR_1 \cap DLD_2$ in the hypergeometric approach, whereas $m_{1:2}$ refers to $(PAR_1 \cap PAR2) \cap DLD_2$ in the DSE based approach. An additional assumption is required in that each draw in DLD_2 is required to be independent, that is the event that one person is captured in DLD_2 has no impact on the probability of any other person from U_2 being captured in DLD_2.

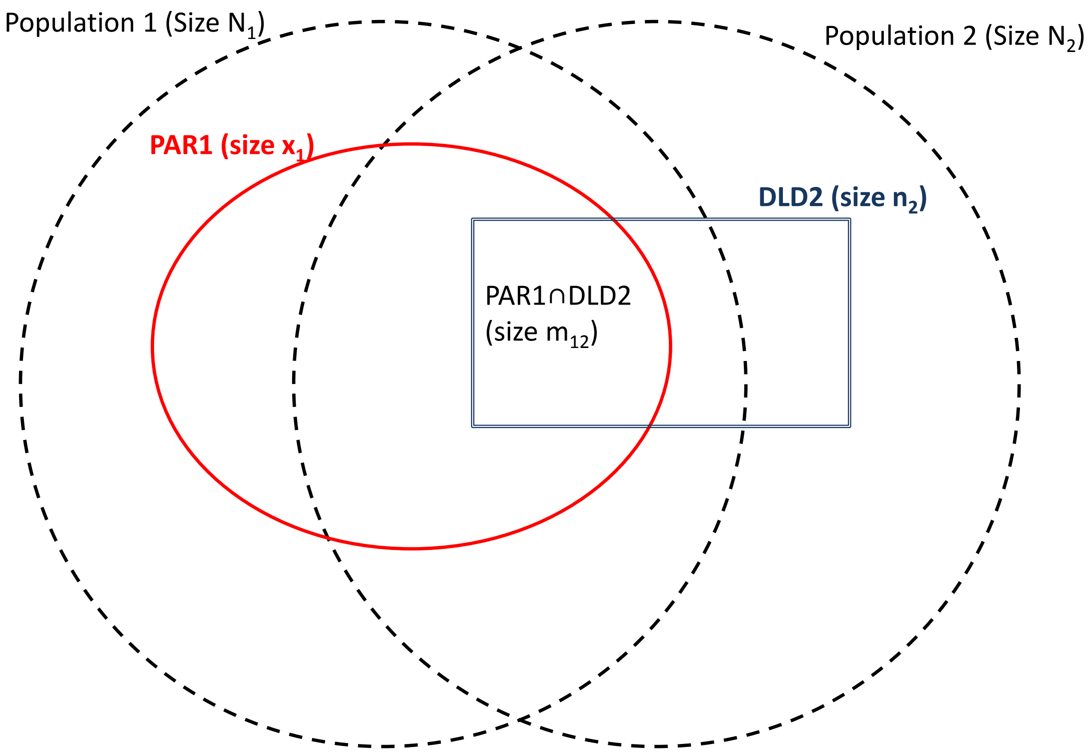

Figure 5.4: Forward based estimate for Stayers using the hypergeometric distribution. Consider the probability of finding m_{12} Stayers in n_2 draws (DLD_2), without replacement, from population 2 of size N_2 that contains exactly K persons that are also in PAR_1.

K_1 is unknown and needs to be estimated. We use $\hat{K}_{1(HYP)}$ to distinguish this estimator from the DSE based estimator $\hat{K}_1$. Fixing N_2 at its estimate $\hat{N}_2 = n_2x_2/m_2$, a Maximum Likelihood Estimate (MLE) $\hat{K}_{1(HYP)}$ is given by the value of K_1 that maximises

$$L(K_1; m_{12}, n_2, N_2) \propto \frac{\binom{K_1}{m_{12}}\binom{N_2-K_1}{n_2-m_{12}}}{\binom{N_2}{n_2}} \tag{5.10}$$

The estimate $\hat{K}_{1(HYP)}$ can be obtained by iterating through the function, equation 5.10, with all possible values of K_1 and selecting the value of K_1 that maximises $L(K_1; m_{12}, n_2, N_2)$.

Noting the general curve of the likelihood function (figure 5.5 with only one maximum for K), Zhang (2009) derives an expression we can use for $\hat{K}_{1(HYP)}$. Incorporating our notation into Zhang's derivation, we obtain an MLE for $K_{1(HYP)}$ as follows:

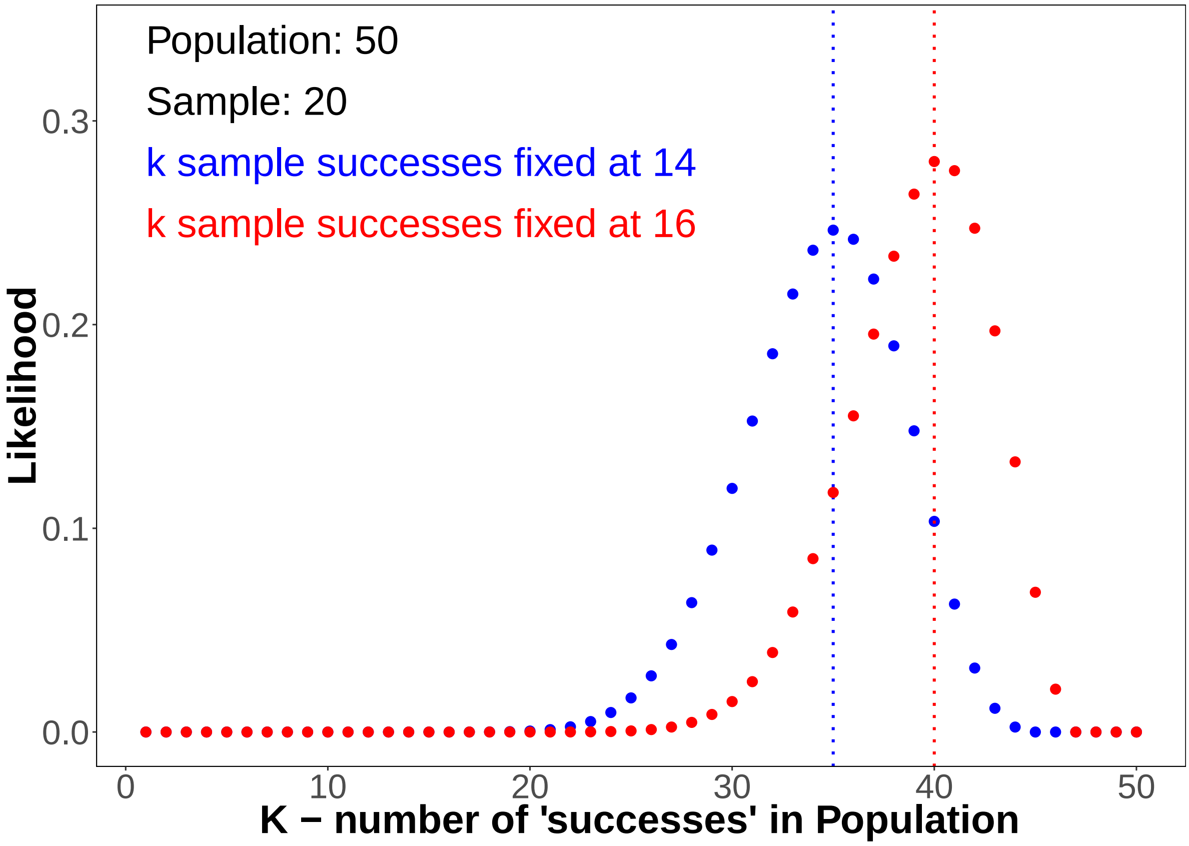

Figure 5.5: Likelihood function plotted for two different scenarios. Each of the two functions plotted has only one maximum

$$\hat{K}_{1(HYP)} = \begin{cases} \frac{m_{12}(N_2+1)}{n_2} - 1 \text{ or } \frac{m_{12}(N_2+1)}{n_2}, & \text{if } \frac{m_{12}(N_2+1)}{n_2} \text{ is an integer} \\ \\ \left\lfloor \frac{m_{12}(N_2+1)}{n_2} \right\rfloor, & \text{if } \frac{m_{12}(N_2+1)}{n_2} \text{ is not an integer} \end{cases} \tag{5.11}$$

where $\lfloor x \rfloor$ denotes the floor of x (or integer part, as we are dealing with positive numbers).

We now derive the variance of the estimator $\hat{K}_{1(HYP)}$.

Using equation 2.4 (page 43) we substitute n_2x_2/m_2 for N_2 in equation 5.11 and write, noting the terms -1 and m_{12}/n_2 have very small impact in the overall picture.

$$\hat{K}_{1(HYP)} \approx \frac{x_2m_{12}}{m_2} + \frac{m_{12}}{n_2}$$

$$\hat{K}_{1(HYP)} \approx \frac{x_2m_{12}}{m_2} \tag{5.12}$$

We also note the form of $V[a/b]$ as follows

$$V\left[\frac{a}{b}\right] \approx \frac{V\left[a\right]}{E\left[b\right]^2} - \frac{2E\left[a\right]}{E\left[b\right]^3}Cov\left[a,b\right] + \frac{E\left[a\right]^2}{E\left[b\right]^4}V\left[b\right] \tag{5.13}$$

and write

$$V\left[\hat{K}_{1(HYP)}\right] \approx x_2^2\left[\frac{V\left[m_{12}\right]}{E\left[m_2\right]^2} - \frac{2E\left[m_{12}\right]}{E\left[m_2\right]^3}Cov\left[m_{12},m_2\right] + \frac{E\left[m_{12}\right]^2}{E\left[m_2\right]^4}V\left[m_2\right]\right] \tag{5.14}$$

We now consider $V[m_2]$. m_2 is the size of the match between PAR and DLD when estimating N_2, where we have assumed independent, homogeneous capture events, and as such, considering x_2 Bernoulli events we can write

$$V\left[m_2\right] = x_2(\pi_2)(1-\pi_2)$$

where an estimate of π_2 can be written as $\hat{\pi}_2 = m_2/x_2$ such that a variance estimator for m_2 can now be written as

$$\hat{V}\left[m_2\right] = x_2\frac{m_2}{x_2}\left[1-\frac{m_2}{x_2}\right]$$

$$= m_2\left[1-\frac{m_2}{x_2}\right] \tag{5.15}$$

We now consider $V[m_{12}]$. We use the Hypergeometric distribution with parameters $N = N_2$, $K = K_1$ and $n = n_2$ to consider the variance as

$$V\left[m_{12}\right] = n_2\left[\frac{K_1}{N_2}\right]\left[1-\frac{K_1}{N_2}\right]\left[\frac{(N_2-n_2)}{(N_2-1)}\right]$$

and noting

$$\frac{\hat{K}_1}{\hat{N}_2} = \frac{x_2 m_{12}}{m_2}\frac{m_2}{x_2 n_2} = \frac{m_{12}}{n_2}$$

and again substituting x_2n_2/m_2 for $\hat{N}_2$ we can write

$$\frac{(\hat{N}_2 - n_2)}{(\hat{N}_2 - 1)} = \frac{(x_2n_2 - m_2n_2)}{(x_2n_2 - m_2)} = \frac{n_2(x_2 - m_2)}{(x_2n_2 - m_2)}$$

we can now write an estimator of the variance of m_{12} as

$$\hat{V}\left[m_{12}\right] = n_2 \left[\frac{m_{12}}{n_2}\right] \left[1 - \frac{m_{12}}{n_2}\right] \left[\frac{n_2(x_2 - m_2)}{(x_2n_2 - m_2)}\right]$$

$$\hat{V}\left[m_{12}\right] = m_{12} \left[1 - \frac{m_{12}}{n_2}\right] \left[\frac{n_2(x_2 - m_2)}{(x_2n_2 - m_2)}\right] \tag{5.16}$$

We now investigate the covariance of m_{12} and m_2. To do this, we partition m_{12} as follows:

$$m_{12} = \sum_{i \in (U_2 \cap PAR_1) \setminus PAR_2} \delta_i + \sum_{i \in (PAR_1 \cap PAR_2)} \delta_i$$

such that the first part of m_{12} refers to Stayers in PAR_1 not contained in PAR_2 that are caught in DLD_2, and the second part of m_{12} refers to those persons that are in both PAR_1 and PAR_2 that are caught in DLD_2.

We also partition m_2 in a similar fashion as follows:

$$m_2 = \sum_{i \in (PAR_2 \setminus PAR_1)} \delta_i + \sum_{i \in (PAR_1 \cap PAR_2)} \delta_i$$

where the first part of m_2 refers to members of PAR_2 not contained in PAR_1 that are caught in DLD_2, and the second part of m_2 refers to those persons that are in both PAR_1 and PAR_2 that are caught in DLD_2.

Here again, $\delta_i = 1$ if $i \in DLD_2$ and $\delta_i = 0$ otherwise. We also note the homogeneous capture assumption with respect to DLD_2 for U_2 and can write $P(\delta_i \in DLD_2) = m_2/x_2$. We also note the independent capture assumption for each person.

We see that the second part in both m_{12} and m_2 is the same, and independent of the respective first parts and as such we can now write

$$Cov\left[m_{12}, m_2\right] = V\left[\sum_{i\in(PAR_1\cap PAR_2)} \delta_i\right]$$

$$= \sum_{i\in(PAR_1\cap PAR_2)} V\left[\delta_i\right]$$

$$\widehat{Cov}\left[m_{12}, m_2\right] = x_{12}\left[\frac{m_2}{x_2}\right]\left[1-\frac{m_2}{x_2}\right] \tag{5.17}$$

To obtain an expression for the variance of $K_{1(HYP)}$ we now substitute from equations 5.16, 5.15 and 5.17 into equation 5.14 to get

$$\begin{aligned}\hat{V}\left[\hat{K}_{1(HYP)}\right] =& x_2^2\Bigg(\frac{m_{12}}{m_2^2}\left[1-\frac{m_{12}}{n_2}\right]\left[\frac{n_2(x_2-m_2)}{(x_2n_2-m_2)}\right] \\ & -\frac{2m_{12}}{m_2^3}x_{12}\left[\frac{m_2}{x_2}\right]\left[1-\frac{m_2}{x_2}\right]+\frac{m_{12}^2}{m_2^4}m_2\left[1-\frac{m_2}{x_2}\right]\Bigg)\end{aligned}$$

which simplifies to

$$\hat{V}\left[\hat{K}_{1(HYP)}\right] = \frac{x_2m_{12}(x_2-m_2)}{m_2^2}\left(x_2\left[\frac{x_2(n_2-m_{12})}{(x_2n_2-m_2)}\right]-\frac{2x_{12}}{x_2}+\frac{m_{12}}{m_2}\right) \tag{5.18}$$

We can now work in the opposite direction to obtain estimators for $K_{2(HYP)}$

$$\hat{K}_{2(HYP)} = \begin{cases} \frac{m_{21}(N_1+1)}{n_1}-1 \text{ or } \frac{m_{21}(N_1+1)}{n_1}, & \text{if } \frac{m_{21}(N_1+1)}{n_1} \text{ is an integer} \\ \\ \left\lfloor\frac{m_{21}(N_1+1)}{n_1}\right\rfloor, & \text{if } \frac{m_{21}(N_1+1)}{n_1} \text{ is not an integer}\end{cases} \tag{5.19}$$

and in a similar manner as equation 5.12, write

$$\hat{K}_{2(HYP)} \approx \frac{x_1 m_{21}}{m_1} \tag{5.20}$$

$$\hat{V}\left[\hat{K}_{2(HYP)}\right] = \frac{x_1 m_{21}(x_1 - m_1)}{m_1^2}\left(x_1\left[\frac{x_1(n_1 - m_{21})}{(x_1 n_1 - m_1)}\right] - \frac{2x_{12}}{x_1} + \frac{m_{21}}{m_1}\right) \tag{5.21}$$

To obtain our estimate of the number of Stayers in the population we simply use equation 5.4 as proposed in section 5.2.3.

5.3.1.2 Comparison with proposed method

The proposed DSE based method uses the overlap between the PAR in consecutive years, and as such, as the PAR goes to full population coverage the variance of this estimator should go to zero. The Hypergeometric based approach does not have this property as the method only relies on 2 lists, PAR and DLD, in different years. So even if the PAR tends to 100% coverage, the Hypergeometric based estimator will only tend to 0% variance if the corresponding DLD tends to 100% coverage. It is not feasible to foresee a situation where every person will have a driving licence and renew it every year, while it is a feasible operational target to strive for 100% coverage in the PAR as the system develops over time. As we saw in table 3.2 (section 3.2), the PAR coverage rate of the population ranged between 89% and 93% from years 2011 to 2016.

It is anticipated, as more data sources are processed and added to the PAR over time, its coverage with respect to the population will also increase. For this reason, the DSE based method will be preferred to the Hypergeometric based method.

5.3.2 Alternative method using PAR from 2 years in a DSE based approach to estimate Stayers

5.3.2.1 Method summary

The proposed method in this section is a DSE system of estimates where the lists correspond to the Stayers in the PAR for years $i = 1, 2$. While we cannot directly observe these lists, we can estimate their size and directly observe the match as $PAR_1 \cap PAR_2$. This DSE system also requires an additional assumption, the assumption of homogeneous capture in at least one of the lists for the population of Stayers ($U_1 \cap U_2$). This assumption only needs to hold with respect to the population of Stayers and allows the populations of Outflows ($U_1 \setminus U_2$) and Inflows ($U_2 \setminus U_1$) to behave differently. This additional assumption is not compatible with our treatment of the PAR as a fixed list however we consider it below .

For this method, we suppose the PAR lists are independent samples where each person in the population has equal probability of being caught in the PAR sample in a given year regardless of whether or not they have they have been caught in a PAR sample for any other year. We make this supposition with reference to both lists. In the context of this estimator we will refer to this assumption as the *independence assumption*. Despite the likely violation or invalidity of this assumption, we consider this estimator because, as the PAR coverage goes to the full population, the bias associated with any violation of this assumption goes to zero.

Figure 5.6 illustrates this DSE system with the combination of the pink and orange shaded areas corresponding to list A, the light blue and orange shaded areas corresponding to list B and the orange shaded area corresponding to the match.

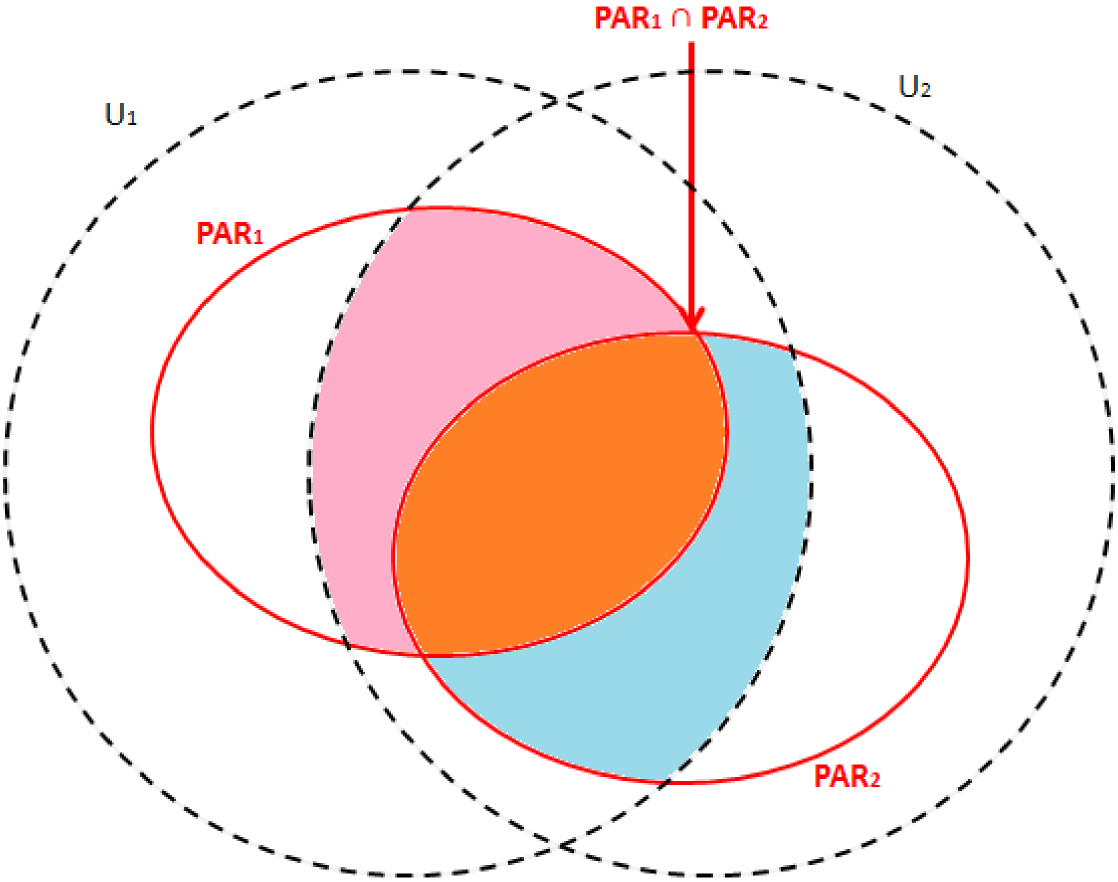

Figure 5.6: Illustration of DSE to estimate Stayers in the Population using estimates of the number of Stayers in PAR_1 and PAR_2

5.3.2.2 Method Development

This method gives the following estimator for the number of Stayers in the population

$$\hat{N}_S = \frac{\hat{K}_1 \hat{K}_2}{x_{1:2}} \tag{5.22}$$

where

$\hat{K}_i$ is the estimated number of Stayers in the SPD (PAR) for year i

$x_{1:2}$ is the number of persons observed as belong to the PAR for both years 1 and 2, and can also be written as $x_{2:1}$. $x_{1:2}$ is constant as it is simply the the overlap of two fixed lists, $PAR_1 \cap PAR_2$.

Now substituting for $\hat{K}_1$ and $\hat{K}_2$ using equations 5.1 and 5.3 respectively, we can rewrite equation 5.22 as

$$\hat{N}_S = \frac{x_{1:2}\frac{n_{1:2}}{m_{1:2}} x_{1:2}\frac{n_{2:1}}{m_{2:1}}}{x_{1:2}}$$

$$= x_{1:2}\frac{n_{1:2}}{m_{1:2}}\frac{n_{2:1}}{m_{2:1}} \tag{5.23}$$

The variance estimator $\hat{V}\left[\hat{N}_S\right]$ now needs to consider the 4 random variables arising from the DLD samples obtained under the homogeneous capture assumption. The 4 random variables are

$$n_{1:2} = \sum_{i \in PAR_1} \delta_{i2}$$

$$m_{1:2} = \sum_{i \in (PAR_1 \cap PAR_2)} \delta_{i2}$$

$$n_{2:1} = \sum_{i \in PAR_2} \delta_{i1}$$

$$m_{2:1} = \sum_{i \in (PAR_1 \cap PAR_2)} \delta_{i1}$$

where $\delta_{i2} = 1$ if individual i belongs to DLD_2 and 0 otherwise, and $\delta_{i1} = 1$ if individual i belongs to DLD_1 and 0 otherwise.

To obtain the variance and covariance terms for the 4 variables, we assume δ_{i2} has independent Bernoulli distribution with probability π_2 for all stayers $i \in PAR_1$ and $\delta_{i2} \equiv 0$ if $i \in PAR_1 \setminus U_2$, and δ_{i1} has independent Bernoulli distribution with probability π_1 for all stayers $i \in PAR_2$ and $\delta_{i1} \equiv 0$ otherwise. Moreover, we assume $Cov(\delta_{i1}, \delta_{i2}) = -\pi_1\pi_2$ for any stayer $i \in PAR_1 \cup PAR_2$, since $\delta_{i1}\delta_{i2} \equiv 0$ as we have assumed no overlap between DLD_1 and DLD_2. It follows that

$$V[n_{1:2}] = K_1\pi_2(1 - \pi_2)$$

$$V[m_{1:2}] = x_{1:2}\pi_2(1 - \pi_2)$$

$$Cov[n_{1:2}, m_{1:2}] = V(m_{1:2})$$

$$V[n_{2:1}] = K_2\pi_1(1 - \pi_1)$$

$$V[m_{2:1}] = x_{1:2}\pi_1(1 - \pi_1)$$

$$Cov[n_{2:1}, m_{2:1}] = V(m_{2:1})$$

$$Cov[n_{1:2}, m_{2:1}] = \sum_{i \in (PAR_1 \cap PAR_2)} Cov[\delta_{i2}, \delta_{i1}] = -x_{1:2}\pi_1\pi_2 = Cov[n_{2:1}, m_{1:2}]$$

$$Cov[n_{1:2}, n_{2:1}) = \sum_{i \in (PAR_1 \cap PAR_2)} Cov[\delta_{i2}, \delta_{i1}] = -x_{1:2}\pi_1\pi_2 = Cov[m_{1:2}, m_{2:1}]$$

where for $Cov[n_{1:2}, n_{2:1}]$ we notice that $PAR_1 \setminus (PAR_1 \cap PAR_2)$ and $PAR_2 \setminus (PAR_1 \cap PAR_2)$ are disjoint.

It is now possible to obtain an approximate variance estimator using the Taylor expansion of ($n_{1:2}$, $m_{1:2}$, $n_{2:1}$, $m_{2:2}$) around their expectations.

5.3.2.3 Comparison with proposed method

If PAR has high coverage of the population for both years then the impact of any violation of the independence assumption should in theory be small.

Given the exposure of this method to likely violations in the independence assumption we prefer the proposed method. In considering the nature of the PAR samples it is reasonable to assume that there exists a positive dependency between samples in consecutive years. If someone is employed or receiving social welfare benefit in one year, they are more likely also to be employed or receiving social welfare benefit in the following year. Similarly, if someone has no connection with official systems in a given year, they are also more likely to have no connection in the following year. Examples of persons that may not engage with public systems include home makers, persons with alternative, foreign or unofficial sources of income.

Note, we should expect some dependence here and we compile a simple sensitivity analysis table to explore the impact which any dependence might have as part of our consideration of an exploration of the different methods using data simulation in section 5.4.3.

5.3.3 Alternative methods where the 2 year population is first estimated

5.3.3.1 Outline: 2 year population based approach

An alternative approach to estimating the flows is to estimate the size of the 2 year population N_{12} first (rather than the Stayers) and then, through differencing with estimates of the population for each year N_1 and N_2, obtain estimates of the Inflows N_I and Outflows N_O.

Now, noting that we can also write $N_{12} = N_O + N_S + N_I$, we can consider an unbiased estimator for N_{12} as the sum of three DSE based estimators, one for each of N_O, N_S and N_I. We express this in equation 5.24

$$\hat{N}_{12(IDEAL)} = \frac{x_O n_O}{m_O} + \frac{x_S n_S}{m_S} + \frac{x_I n_I}{m_I} \tag{5.24}$$

where

x_O, x_S and x_I refer to components of $PAR_1 \cup PAR_2$ that belong to the Outflows, Stayers and Inflows sub-populations, respectively,

n_O, n_S and n_I refer to components of $DLD_1 \cup DLD_2$ that belong to the Outflows, Stayers and Inflows sub-populations, respectively, and

m_O, m_S and m_I refer to components of $(PAR_1 \cup PAR_2) \cap (DLD_1 \cup DLD_2)$ that belong to the Outflows, Stayers and Inflows sub-populations, respectively.

We refer to this estimator as the IDEAL estimator as it is not possible to directly observe each of the counts, but if we could, it would be a strong candidate for the preferred method. Figure 5.7 illustrates the components of the ideal estimator with the 3 DSEs for Outflows, Stayers and Inflows shaded Orange, Blue and Green. Lighter shades (along with darker shades) denote the components that belong to the SPD, medium shades (along with darker shades) denote the components that belong to *DLD* and the darker shades denote the match used in the DSE.

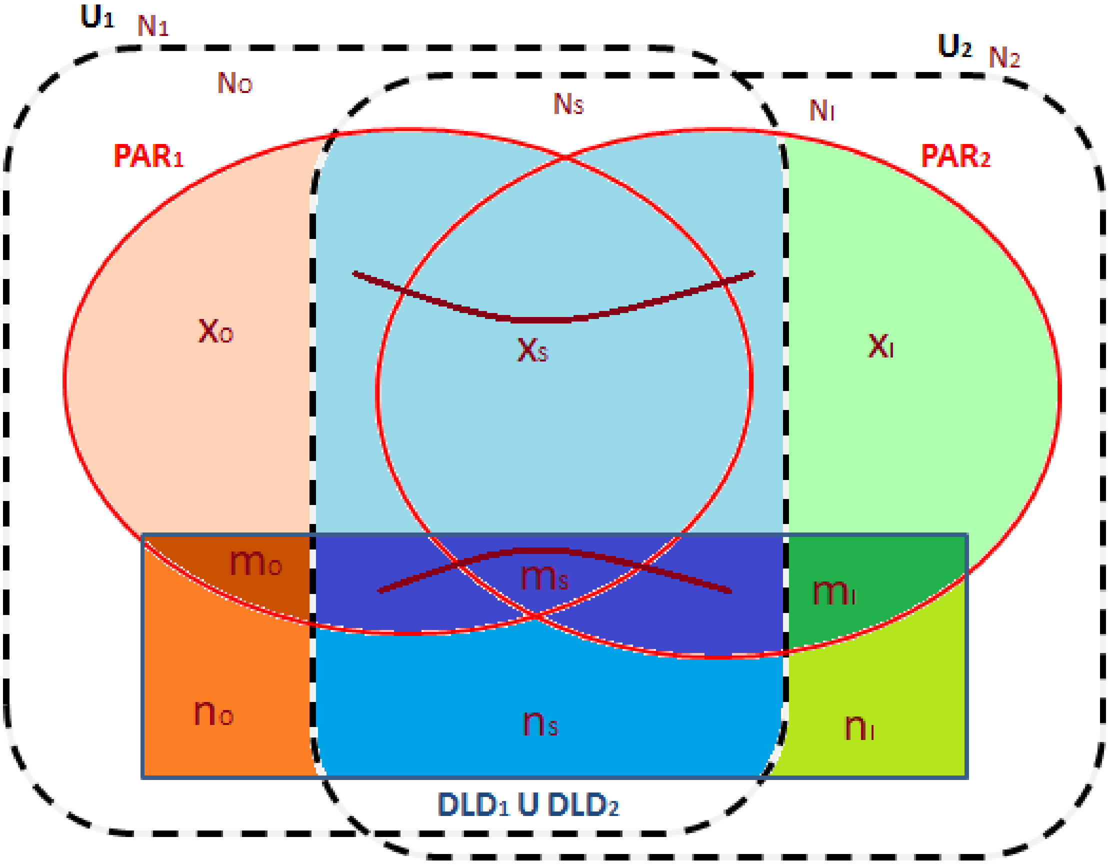

Figure 5.7: Illustration of Ideal DSE to estimate the 2 year population N_{12}, the size of $U_1 \cup U_2$, with 3 separate DSEs for Outflows, Stayers and Inflows shaded Orange, Blue and Green respectively. Quantities for DSE formulas labelled in wine.

If we suppose that the population sizes for Inflows, N_I, and Outflows, N_O, are insignificant in the overall size of the 2 year population, N_{12}, and consider a simpler alternative DSE based estimator as in equation 5.25.

$$\hat{N}_{12(NAIVE)} = \frac{(x_O + x_S + x_I)(n_O + n_S + n_I)}{(m_O + m_S + m_I)} \tag{5.25}$$

By definition, there will be bias in the estimator $\hat{N}_{12.NAIVE}$ as list B $(DLD_1 \cup DLD_2)$ does not satisfy the homogeneous capture assumption. The Outflows part of the 2 year population only has a chance of being caught in DLD_1, the Inflows part of the 2 year population only has a chance to be caught in DLD_2, while the Stayers part of the 2 year population has a significantly higher likelihood of being caught in list B, as it may be caught in DLD_1 in year 1 and it may be caught in DLD_2 in year 2. However, as suggested earlier, the bias in the estimator with respect to violation of the *homogeneous capture* assumption may be small relative to estimating the 2 year population and may be small enough such that it can be ignored (or estimated) when considering the Inflows and Outflows. We will also consider the possibility of adjusting this estimator to eliminate any bias.

The remainder of this section focuses on describing the methodology to estimate the 2 year population. We will first consider the *IDEAL* estimator $N_{12(IDEAL)}$, as given in equation 5.24, before considering a simpler, more naive and easier estimator $N_{12.NAIVE}$ in equation 5.25. The section will also consider the bias in $N_{12.NAIVE}$ by comparing it with $N_{12.IDEAL}$ to see if the estimator can be used or adjusted in some practical way.

5.3.3.2 Consideration of the *IDEAL* 2 year population estimate

We consider each of the terms in equation 5.24, starting with the terms relating to the SPD, x_O, x_S and x_I. We cannot directly count or obtain the values of these terms. However, we can observe or count the persons belonging to $PAR_1 \cup PAR_2$, which we will denote as x_{12}. Note the absence of the colon in the notation here; we used $x_{1:2}$ earlier to denote $PAR_1 \cap PAR_2$ and we will remain consistent with this notation. Using results from section 5.2, we can obtain an estimate of x_S as

$$\hat{x}_S = \hat{K}_1 + \hat{K}_2 - x_{1:2} \tag{5.26}$$

where $\hat{K}_i$ is the number of Stayers in the SPD for year i, $i = 1, 2$ and

$x_{1:2}$ is the size of the population observed or counted in $PAR_1 \cap PAR_2$.

We can now derive estimates for x_O and x_I as follows:

$$\hat{x}_O = x_1 - \hat{x}_S$$

and

$$\hat{x}_I = x_2 - \hat{x}_S$$

where x_1 and x_2 are the observed sizes of the SPD in years 1 and 2.

We can also obtain estimates of m_O and m_I by noting DLD_1 and DLD_2 satisfy the *homogeneous capture assumption* and simply applying the estimated catch rates for the respective populations as follows

$$\hat{m}_O = \hat{x}_O \frac{n_1}{\hat{N}_1}$$

and

$$\hat{m}_I = \hat{x}_I \frac{n_2}{\hat{N}_2}$$

We can also obtain an estimate for m_S as follows

$$\hat{m}_S = m_{12} - \hat{m}_O - \hat{m}_I$$

where m_{12} can be directly observed/counted as $(PAR_1 \cup PAR_2) \cap (DLD_1 \cup DLD_2)$.

The challenge now arises in estimating n_O, n_S and n_I, and these cannot be estimated without first estimating N_S. But, if we can estimate N_S then we already have estimates of N_I and N_O, as described in section 5.2, and, as such, this approach may or may not provide any added value.

5.3.3.3 Consideration of the *NAIVE* 2 year population estimate

From equation 5.25, the estimator and an estimate of its variance is easily obtained, as it is treated as a simple DSE with list A as $x_{12} = x_O + x_S + x_I$, list B as $n_{12} = n_O + n_S + n_I$ and the match as $m_{12} = m_O + m_S + m_I$, with each of the 3 terms in the DSE being directly observable.

While the advantage of this estimator is that it is easy to calculate, the drawback is that there will be bias due to a violation of the *homogeneous capture assumption.*

We will now set out to estimate this bias and, where necessary, find a mechanism for adjusting this estimator for bias. For simplicity in notation we will now use $\hat{N}$ to denote the naive estimator $\hat{N}_{12.NAIVE}$ and $\tilde{N}$ to denote the ideal estimator $\hat{N}_{12.IDEAL}$. We restate the estimators with the new notation.

$$\tilde{N} = \frac{x_O n_O}{m_O} + \frac{x_S n_S}{m_S} + \frac{x_I n_I}{m_I}$$

giving an expected value

$$E\left[\tilde{N}\right] = N_O + N_S + N_I$$

$$\hat{N} = \frac{(x_O + x_S + x_I)(n_O + n_S + n_I)}{(m_O + m_S + m_I)}$$

We know

$$E\left[\frac{m_O}{x_O}\right] = E\left[\frac{n_1}{N_1}\right] = \pi_O$$

$$E\left[\frac{m_I}{x_I}\right] = E\left[\frac{n_2}{N_2}\right] = \pi_I$$

$$E\left[\frac{m_S}{x_S}\right] = E\left[\frac{n_1}{N_1} + \frac{n_2}{N_2}\right] = \pi_O + \pi_I$$

and use $E\left[m_O\right] = \pi_O x_O$ to write the expected value of $\hat{N}$ as

$$E\left[\hat{N}\right] = \frac{(x_O + x_S + x_I)E\left[n_O + n_S + n_I\right]}{E\left[m_O + m_S + m_I\right]}$$

$$= \frac{(x_O + x_S + x_I)E\left[n_O + n_S + n_I\right]}{(\pi_O x_O + \pi_O x_S + \pi_I x_S + \pi_I x_I)}$$

$$= \frac{(x_O + x_S + x_I)\left[\pi_O(N_O + N_S) + \pi_I(N_S + N_I)\right]}{(\pi_O x_O + \pi_O x_S + \pi_I x_S + \pi_I x_I)} \tag{5.27}$$

In order, to simplify the algebraic workings, we now assume some stability in the catch rate. That is, we assume $E[n_1/N_1] = E[n_2/N_2]$ and hence $\pi_O = \pi_I = \pi$, and write

$$E\left[\hat{N}\right] \approx \frac{(x_O + x_S + x_I)\pi(N_O + N_S + N_I + N_S)}{\pi(x_O + x_S + x_O + x_S)}$$

$$= \frac{(x_O + x_S + x_I)(N_O + N_S + N_I + N_S)}{(x_O + x_S + x_I + x_S)} \tag{5.28}$$

Next, we divide above and below by $(x_O + x_S + x_I)$ to get

$$E\left[\hat{N}\right] \approx \frac{(N_O + N_S + N_I + N_S)}{1 + \dfrac{x_S}{(x_O + x_S + x_I)}} \tag{5.29}$$

Now, we divide above and below by $(N_O + N_S + N_I)$ to get

$$E\left[\hat{N}\right] \approx (N_O + N_S + N_I)\frac{1 + \dfrac{N_S}{(N_O + N_S + N_I)}}{1 + \dfrac{x_S}{(x_O + x_S + x_I)}} \tag{5.30}$$

now, if we write

$$\xi_N = \frac{N_S}{(N_O + N_S + N_I)}$$

and

$$\xi_x = \frac{x_S}{(x_O + x_S + x_I)}$$

we get

$$E\left[\hat{N}\right] \approx \tilde{N}\frac{1 + \xi_N}{1 + \xi_x}$$

and we can observe that the direction of the bias is dependent on the ratio of the proportion of Stayers in the 2 year population over the proportion of Stayers captured in the SPD over two years. That is, if $\xi_N > \xi_x$ then $E\left[\hat{N}\right] > \tilde{N}$, if $\xi_N < \xi_x$ then $E\left[\hat{N}\right] < \tilde{N}$ and if $\xi_N = \xi_x$ then $E\left[\hat{N}\right] = \tilde{N}$.

We can estimate ξ_x as

$$\hat{\xi}_x = \frac{\hat{x}_S}{x_{12}} \tag{5.31}$$

where $\hat{x}_S$ is estimated using equation 5.26 and x_{12} is directly observed as $PAR_1 \cup PAR_2$.

It is also possible to estimate ξ_N using either of equations 5.4 or 5.22 to estimate N_S and noting the additional assumptions to be made. We note also that the size of the PAR is close to the population size in real life (usually having greater than 90% coverage) and, as such, the impact of any weakness or violation in assuming the PAR is representative of the population may only have a small effect.

We estimate ξ_N as

$$\hat{\xi}_N = \frac{\hat{N}_S}{\hat{N}_1 + \hat{N}_2 - \hat{N}_S} \tag{5.32}$$

This now allows us the possibility of creating a new adjusted version of the naive estimator $\hat{N}$, which we will call $\hat{N}_{(ADJ)}$ and we will estimate as follows:

$$\hat{N}_{(ADJ)} = \frac{(1+\hat{\xi}_x)}{(1+\hat{\xi}_N)}\hat{N} \tag{5.33}$$

which should contain less bias than $\hat{N}$.

Given the complexity of this expression we will rely on bootstrap methods to estimate the variance of this estimator.

5.3.3.4 Comparison of 2 year population based approach with proposed method

It is not feasible to implement the *IDEAL* 2 year population based estimator so this is not a candidate.

The *NAIVE* 2 year population based estimator will not be preferred to the proposed estimator as there are obvious weaknesses in the underlying assumptions. The impact of these weaknesses will be explored later using data simulations.

Adjusting the *NAIVE* 2 year population based estimator will provide some improvement over the naive method, however if the adjustment is based on using the proposed method to estimate Stayers it is difficult to see what possible improvements this method will have over the proposed method.

5.3.4 Summary of proposed and alternative estimators

In summary, we have proposed a method to estimate population flows which we label $M1_{(DSE)}$. We also propose a number of alternatives, the first being a variant on the proposed method which we label $M1_{(HYP)}$. The M1 in the label denotes it being a similar class of estimate, in that it first estimates the Stayers in the PAR before grossing up to the population. The labeling of the estimators then follows this sequence with the M label denoting the *class* of estimator and the subscript denoting the variant (typically whether the estimator relies on using DSE based methods or Hypergeometric based methods to estimate the Stayers in the PAR.

The M1 and M2 class of estimators are based on estimating the Stayers in the population (N_S) and then differencing with the estimated population sizes $\hat{N}_1$ and $\hat{N}_2$ to obtain estimates of Outflows ($\hat{N}_O$) and Inflows ($\hat{N}_I$). Both classes also rely first on estimating the number of Stayers in the SPD before estimating the number of Stayers in the population. The M1 class of estimator assumes the SPD is representative of the population and grosses up an estimate of Stayers in the SPD. The M2 class of estimator relies on a DSE derived from the SPDs to estimate the Stayers in the population, where the two list sizes are given by the estimated number of Stayers in the respective SPD.

The primary difference between these two classes lies in the underlying assumptions. For the M1 class of estimator, there is an an extra assumption in that the SPD is considered representative of the population as a whole, while for the M2 class the assumption is much weaker in that it only requires the Stayers cohort of the population to satisfy the *homogeneous capture assumption* with respect to one of the SPDs.

Both classes of estimator rely first on estimating the number of Stayers in the SPD, for which two methods have been identified, one based on a DSE approach and the other based on an application of the hypergeometric distribution. Both applications use information that the other application doesn't and as such there may be scope to combine both of these approaches into a hybrid one, however as the PAR goes to full population coverage, the variance of the DSE based approach goes to zero and as such combining with the hypergeomteric based approach is unlikely to provide much benefit.

The M3, M4 and M5 classes of estimator rely first on estimating the *2 year population* before differencing between the population sizes in year 1 and 2 to obtain estimates of Outflows ($\hat{N}_O$) and Inflows ($\hat{N}_I$). The M3 class of estimator does not rely on first estimating the number of Stayers in the SPD, but simply makes an assumption that, given the size of the gross flows is so small in comparison to the population, the impact of any violation of the *homogeneous capture assumption* in the underlying DSE model will be negligible. This assumption does not hold up, however this class of estimator is then the basis for the M4 and M5 class of estimator. The M4 and M5 classes of estimator simply take the M3 class of estimator and adjust it for the violation of the *homogeneous capture assumption* using the first and second estimators, respectively.

We will use M1 to M5 to denote these estimators in the simulation exercises in the next section. We will also add a sixth estimator (denoted by M6), based on knowing which persons in the samples belong to the underlying populations of Outflows (U_O), Inflows (U_I), and Stayers (U_S). It is not possible to calculate M6 in the real world but it is possible in the simulation exercises. We call M6 the Ideal estimator.

5.4 Simulation exercise

5.4.1 Overview of Simulations

The following describes the simulation exercises.

The first step is the creation of a simulated population of Stayers (U_S), Inflows (U_I) and Outflows (U_O) of known sizes N_S, N_I and N_O respectively. In the real world, these population sizes are unknown. The second step is then to draw the 2 PAR samples PAR_1 and PAR_2 and hold them fixed for the methods where they are assumed to be fixed (M1, M3 and M4). For each iteration of the simulation, the samples will be

Method	Description
$M1_{(DSE)}$	**Proposed**. Estimation based on estimating number of Stayers based in population by first estimating number of Stayers in each of the PAR's using DSE methods and then scaling this estimate up to the population based on the ratio of the population size to the size of PAR.
$M1_{(HYP)}$	Estimation based on estimating number of Stayers based in population by first estimating number of Stayers in each of the PAR's using an application of the hypergeometric distribution and then scaling this estimate up to the population based on the ratio of the population size to the size of PAR.
$M2_{(DSE)}$	Based on estimating the number of Stayers first. Estimation based on DSE methods where the size of list A is based on an estimate of the number of Stayers in PAR_1 and the size of list B is based on an estimate of the number of Stayers in PAR_2. The estimates of both lists are obtained using DSE methods.
$M2_{(HYP)}$	Based on estimating the number of Stayers first. Estimation based on DSE methods where the size of list A is based on an estimate of the number of Stayers in PAR_1 and the size of list B is based on an estimate of the number of Stayers in PAR_2. The estimates of both lists are obtained using using applications of the Hypergeometric distribution.
$M3_{(NAIVE)}$	A simple DSE based approach to estimating the *2 year population* first. List A is based on the union $PAR_1 \cup PAR_2$ and list B is based on the union $DLD_1 \cup DLD_2$
$M4_{(DSE)}$	Adjusts for the bias in $M3_{(NAIVE)}$ using estimates from $M1_{(DSE)}$
$M4_{(HYP)}$	Adjusts for the bias in $M3_{(NAIVE)}$ using estimates from $M1_{(HYP)}$
$M5_{(DSE)}$	Adjusts for the bias in $M3_{(NAIVE)}$ using estimates from $M2_{(DSE)}$
$M5_{(HYP)}$	Adjusts for the bias in $M3_{(NAIVE)}$ using estimates from $M2_{(HYP)}$
$M6_{(IDEAL)}$	Based on the sum of 3 estimates for Inputs, Outputs and Stayers obtained using DSE methods. Not possible in real world applications as we cannot determine this partition into DLD_1 and DLD_2. However, it is available in simulation exercises and, as such, provides a benchmark for the other estimators.

Table 5.1: A summary of the proposed and alternative methods to estimate Gross Population Flows.

	In PAR	Not in PAR	
In Stayers	m_{11}	m_{12}	m_{1+}
Not in Stayers	m_{21}	m_{22}	m_{2+}
	m_{+1}	m_{+2}	m_{++}

Table 5.2: Table used to estimate population totals broken down by whether a person is included in the PAR and whether a person belongs to the population of Stayers, based on a given odds ratio $\theta = (m_{11}m_{22})/(m_{12}m_{21})$

drawn according to the sampling environment that is being simulated. The different estimators will then be compiled for each iteration to give a dataset containing a row for each iteration. This resulting dataset can then be used to evaluate the performance of the different estimators.

The DLD simulated samples will take the form of a simple random sample across the relevant parts of the population, U_O and U_S for DLD_1 and U_S and U_I for DLD_2.

Odds ratios are used to simulate PAR samples with a specified likelihood of belonging to the population of Stayers. The odds ratio (θ) can be interpreted as the odds of a person belonging to the population in two consecutive time periods (Stayers), given the knowledge of whether a person is a member in the PAR for the time period in question, compared to those same odds, given no knowledge of whether a person belongs in the PAR. Using table 5.2 to represent counts belonging to the relevant PAR and Stayers groups, the odds ratio can be expressed as

$$\theta = \frac{m_{11} \,/\, m_{21}}{m_{12} \,/\, m_{22}}$$

or

$$\theta = \frac{m_{11}m_{22}}{m_{12}m_{21}}$$

$\theta = 1$ is the special case where knowing a person is in the PAR is of no added value to determining if that person is a member of U_S. That is, every person in the population for year 1, ($U_1 = U_O \cup U_S$), has an equal probability of being included in PAR_1. Similarly, every person in the population for year 2 has an equal probability of being included in PAR_2.

The marginals are the population characteristics and, for the purpose of the simulation exercise, are known as follows:

m_{++} is the size of the population in year i (N_i)

m_{1+} is the number of Stayers (S) in the population (N_S)

m_{2+} is the number of Non Stayers $N_i - N_S$ in the population (can be inflows or outflows depending on whether looking at population in year 1 or year 2).

m_{+1} is the number of the population in year i in the PAR_i (x_i)

m_{2+} is the number of the population in year i not in the PAR_i $(N_i - x_i)$

N and x can refer to the population and PAR in either the first or second time period, when considering the two time periods for where Stayers belong to the population. Using the relationships ($m_{11} + m_{12} = m_{1+}$, $m_{21} + m_{22} = m_{2+}$, $m_{11} + m_{21} = m_{+1}$ and $m_{12} + m_{22} = m_{+2}$) from table 5.2, and expressing the odds ratio relationship in terms of m_{11} before then solving for m_{11}, will allow us to determine the expected value of the table cells in table 5.2 for any given value of θ. Expressing the odds ratio in terms of m_{11} gives an equation in quadratic form to be solved for m_{11} as follows

$$am_{11}^2 + bm_{11} + c = 0$$

giving

$$m_{11} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where

$$\begin{aligned} a =& 1 - \theta \\ b =& m_{2+} + (\theta - 1)m_{+1} + \theta m_{1+} \\ c =& -\theta m_{+1}m_{1+} \end{aligned}$$

Once m_{11} is known, the other cell counts are obtained by differencing from the margins using table 5.2.

A PAR sample of size x_i with a specific odds ratio can now be sampled from a given population as follows:

A draw from the binomial distribution with parameters ($n = N_S$, $p = m_{11}/N_S$) is used to simulate the number of persons x_S to sample from U_S. Samples of sizes x_S and $x_i - x_S$ are drawn from U_S and $U_i \setminus U_S$, respectively, using Simple Random Sampling and combined to give the simulated PAR sample of x_i for year i.

Setting up the simulation to create the PAR samples in this way allows the exploration of representativity in terms of Stayer/Non Stayer split using different values of θ. For example, values of 0.5 and 2 for θ relate to the PAR being half as likely and twice as likely to contain Stayers than the population not covered by the PAR.

5.4.2 Simulation - Significant population size and high PAR coverage

The population size parameters for the simulation are chosen to reflect young Irish adult population groups to see what level of precision and accuracy may be expected if these estimators were to be deployed. The configuration of the simulation can be adjusted to represent any population groupings. As such, it can also be used to obtain a bootstrap estimate of the variance of the estimators as is done later.

The underlying population used in the simulation consists of 30,000 persons belonging to U_S (Stayers) and 1,000 persons belong to each of U_I and U_O (Inflows and Outflows). The PAR has a sample size of 27,000 in each year and the DLD has a sample size of 3,000 in each year. The simulation with different values of θ is applied to the PAR sample selection. θ provides a measure of the representativity of the PAR samples with respect to the population of Stayers.

We will use the Relative Root Mean Squared Error (RRMSE) as the overall measure of how well each estimator performs, with the Relative Bias (RBias) and Coefficient of Variation (CV) providing measures of accuracy and precision, respectively. We consider estimator $M6$ as the gold standard. This is the best we could do if it was possible to determine for each person in the samples if they belonged to the population of Stayers or not. The performance measures for the simulations are provided in tables 5.3, 5.4 and 5.5 for values of $\theta = 1, 2, 0.5$ respectively.

In considering the general performance of the estimators, the RRMSE tends to be between 24% and 30%. $M2_{(HYP)}$ is the exception, tending to be higher with scores just over 40%. In the context of estimating population flows of the order of 1 in 31 persons the performance measures indicate that there is some value to these estimators.

In considering whether there is any value in using the hypergeometric approach to estimate Stayers in the PAR, we can see for each of the estimators that the DSE based versions perform better. We note the simulation exercises have been set up such that the PAR has a high coverage rate (90%), while the DLD has a relatively low coverage rate (10%).

Bias tends to be very small in comparison to standard deviation for each of the measures regardless of whether PAR is representative of the population or not.

In considering the sensitivities of the different estimators to whether or not PAR is representative of the population with respect to Stayers, $M2$ and $M5$ are less susceptible to bias than $M1$ and $M4$. $M1$ and $M4$ both assume that PAR samples are representative of the population. However, we note from the simulation that any bias introduced due

Estimator	mean	sd	Bias	RBias (%)	CV (%)	RMSE	RRMSE (%)
$M1_{(DSE)}$	1008.0	221.2	8.0	0.8	21.9	221.4	22.1
$M1_{(HYP)}$	1005.7	285.8	5.7	0.6	28.4	285.9	28.6
$M2_{(DSE)}$	1010.2	246.6	10.2	1.0	24.4	246.8	24.7
$M2_{(HYP)}$	1005.7	419.3	5.7	0.6	41.7	419.3	41.9
$M3_{(NAIVE)}$	896.6	212.8	-103.4	-10.3	23.7	236.6	23.7
$M4_{(DSE)}$	1007.9	222.2	7.9	0.8	22.0	222.3	22.2
$M4_{(HYP)}$	1006.8	224.1	6.8	0.7	22.3	224.2	22.4
$M5_{(DSE)}$	1010.1	245.5	10.1	1.0	24.3	245.7	24.6
$M5_{(HYP)}$	1006.0	313.6	6.0	0.6	31.2	313.6	31.4
$M6_{(IDEAL)}$	1000.5	40.2	0.5	0.0	4.0	40.2	4.0

Table 5.3: Simulated results for inflows. Population Stayers=30,000; Inflow=1,000; Outflow=1,000 . Sampling $PAR_1 = PAR_2 = 27,000$; $DLD_1 = DLD_2 = 3,000$; Simulations=1,000 . PAR samples selected as simple random samples from respective populations. M1 and M4 class of estimators are based on averaging both ways (using DLD_1 and DLD_2) of estimating Stayers and then obtaining the empirical variance estimate. ($\theta = 1$)

Estimator	mean	sd	Bias	RBias (%)	CV (%)	RMSE	RRMSE (%)
$M1_{(DSE)}$	885.4	232.5	-114.6	-11.5	26.3	259.2	25.9
$M1_{(HYP)}$	882.0	287.9	-118.0	-11.8	32.6	311.1	31.1
$M2_{(DSE)}$	994.5	253.8	-5.5	-0.6	25.5	253.8	25.4
$M2_{(HYP)}$	987.7	409.1	-12.3	-1.2	41.4	409.3	40.9
$M3_{(NAIVE)}$	788.0	223.0	-212.0	-21.2	28.3	307.7	30.8
$M4_{(DSE)}$	885.2	232.9	-114.8	-11.5	26.3	259.7	26.0
$M4_{(HYP)}$	884.1	234.4	-115.9	-11.6	26.5	261.5	26.1
$M5_{(DSE)}$	994.3	252.4	-5.7	-0.6	25.4	252.5	25.2
$M5_{(HYP)}$	989.1	310.6	-10.9	-1.1	31.4	310.8	31.1
$M6_{(IDEAL)}$	1001.8	58.8	1.8	0.2	5.9	58.8	5.9

Table 5.4: Simulated results for inflows. Population Stayers=30,000; Inflow=1,000; Outflow=1,000 . Sampling $PAR_1 = PAR_2 = 27,000$; $DLD_1 = DLD_2 = 3,000$; Simulations=1,000 . Persons selected into the PAR sample are twice as likely to belong to the population of Stayers as those not selected into the PAR sample. M1 and M4 class of estimators are based on averaging both ways (using DLD_1 and DLD_2) of estimating Stayers and then obtaining the empirical variance estimate. ($\theta = 2$)

Estimator	mean	sd	Bias	RBias (%)	CV (%)	RMSE	RRMSE (%)
$M1_{(DSE)}$	1062.1	224.9	62.1	6.2	21.2	233.3	23.3
$M1_{(HYP)}$	1068.5	283.7	68.5	6.8	26.5	291.8	29.2
$M2_{(DSE)}$	995.6	245.9	-4.4	-0.4	24.7	245.9	24.6
$M2_{(HYP)}$	1008.4	412.7	8.4	0.8	40.9	412.8	41.3
$M3_{(NAIVE)}$	943.4	214.6	-56.6	-5.7	22.7	221.9	22.2
$M4_{(DSE)}$	1063.0	224.9	63.0	6.3	21.2	233.6	23.4
$M4_{(HYP)}$	1062.6	226.4	62.6	6.3	21.3	234.9	23.5
$M5_{(DSE)}$	996.5	243.7	-3.5	-0.4	24.5	243.8	24.4
$M5_{(HYP)}$	1001.8	307.3	1.8	0.2	30.7	307.3	30.7
$M6_{(IDEAL)}$	1000.2	30.4	0.2	0.0	3.0	30.4	3.0

Table 5.5: Simulated results for inflows. Population Stayers=30,000; Inflow=1,000; Outflow=1,000 . Sampling $PAR_1 = PAR_2 = 27,000$; $DLD_1 = DLD_2 = 3,000$; Simulations=1,000 . Persons selected into the PAR sample are half as likely to belong to the population of Stayers as those not selected into the PAR sample. M1 and M4 class of estimators are based on averaging both ways (using DLD_1 and DLD_2) of estimating Stayers and then obtaining the empirical variance estimate. ($\theta = 0.5$)

to violation of this representativity assumption tends to be very small in comparison to the standard deviation of the estimates. We also note here that $M2$ and $M5$ rely on an added or different assumption, that there is no dependence between Stayers in PAR_1 and PAR_2, and we will look at the sensitivity of the estimators with respect to violation of this assumption in the next section.

Population flows are a very small part of the population and extremely difficult to pick up unless there are formal systems in place at borders to record each person as they enter and exit the country. Given the performance measures presented in this section, each of the estimators has significant value in providing a usable estimate of the population flows.

We also take the opportunity to compare the simulated variance for N_S under the proposed method ($M1_{(DSE)}$) with the estimated variance from equation 5.5 when $\theta = 1$. The simulated variance of N_S is 93,025 while the estimated variance is 98,523 with an empirical 95% confidence interval of [92,320 - 104,430] (based on 1,000 simulations). The ratio of the square roots of the estimated to the simulated variance is 1.06 indicating a 6% difference between CVs. We would expect the simulated variance to be less than the estimated variance for 2 reasons. The first reason is that the estimated variance is considered a conservative estimate with a covariance term being ignored. The second reason is that, in practice, where an unrealistic estimate is obtained outside the known bounds the estimate is truncated to the nearest bound. The estimated number of Stayers cannot be greater than the minimum population size in years 1 and 2 (in practice the population sizes are also estimated) and the estimated number of Stayers cannot be less

	In PAR_1	Not in PAR_1	
In PAR_2	m_{11}	m_{12}	m_{1+}
Not in PAR_2	m_{21}	m_{22}	m_{2+}
	m_{+1}	m_{+2}	m_{++}

Table 5.6: Table used to estimate cell totals with respect to the population broken down by whether a person is included in PAR in Year 1 and whether a person is included in PAR in year 2, based on a given odds ratio $\theta = (m_{11}.m_{22})/(m_{12}.m_{21})$

θ	m_{11}	m_{12}	m_{21}	m_{22}	$\hat{N}_S$	Bias	% Bias	% Bias
						($\hat{N}_S$)	($\hat{N}_S$)	($\hat{N}_I$)
0.10	24,037	2,963	2,963	37	30,328	328	1.1	-32.8
0.20	24,071	2,929	2,929	71	30,285	285	1.0	-28.5
0.50	24,166	2,834	2,834	166	30,166	166	0.6	-16.6
1.00	24,300	2,700	2,700	300	30,000	0	0	0
2.00	24,507	2,493	2,493	507	29,747	-253	-0.8	25.3
5.00	24,892	2,108	2,108	892	29,287	-713	-2.4	71.3
10.00	25,235	1,765	1,765	1,235	28,888	-1,112	-3.7	111.2

Table 5.7: Sensitivity Analysis with respect to dependencies between PAR samples when estimating Stayers in the population using DSE (M2 class). $N_S = 30,000$; $x_1 = x_2 = 27,000$; $N_I = N_O = 1,000$. $\theta = 2$ means that persons selected into the first PAR sample are twice as likely to be selected in the second PAR sample compared to those that were not selected in the first PAR sample.

than $\#(PAR_1 \cup DLD_1) \cap (PAR_2 \cup DLD_2)$ the number of Stayers identified by comparing the 4 lists. The simulations incorporate this truncation.

5.4.3 Dependency between PAR samples - sensitivity analysis for estimate of Stayers (M2)

We revisit the odds ratio development in section 5.4.1 to develop a sensitivity analysis for estimates of Stayers and Inflows when using PAR samples and DSE (M2 class). We begin by restating table 5.2 in terms of PAR_1 and PAR_2 in table 5.6. $\theta = (m_{11}m_{22})/(m_{12}m_{21})$ now compares the odds of a person being included in PAR_2 when they are included in PAR_1 with the odds of a person being included in PAR_2 when they are not included in PAR_1 and expresses this comparison as a ratio. When $\theta = 1$ there is no dependency between the PAR samples and, as such, the DSE estimator (M2) will produce an unbiased estimator of N_S.

We use the workings in section 5.4.1 to obtain an expected value of m_{11} under different values of θ and then derive values for m_{12} and m_{21} knowing the marginal totals (size of PAR samples), before estimating $\hat{N}_S = m_{1+}m_{+1}/m_{11}$. The same population and sample size values used for the data simulations in section 5.4.2 are used again here and the results are shown in table 5.7.

In considering the nature of the PAR samples it is reasonable to assume that there exists a positive dependency between samples in consecutive years. If someone is employed or receiving social welfare benefit in one year, they are more likely to also be employed or receiving social welfare benefit in the following year. Similarly, if someone has no connection with official systems in a given year, they are also more likely to have no connection in the following year. Examples of persons that may not engage with public systems include home makers, persons with alternative, foreign or unofficial sources of income. For this reason, we choose wide ranging values of θ from 0.1 to 10. Positive values of θ signify a positive dependence between the two PAR samples.

From table 5.7 we can see that the impact on the estimate of Stayers is relatively small for the different dependency values (θ's) chosen, with the bias ranging from 1.1% when $\theta = 0.1$ (negative dependence) to -3.7% when $\theta = 10$ (positive dependence). The bias is small in comparison to the estimate of Stayers and this can be expected, given that the coverage of both PAR samples is high (90%). However, if we consider the impact of this bias on the estimate of Inflows ($\hat{N}_I$), which is very small relative to the estimate of Stayers ($\hat{N}_S$), we have a different story. When we consider the value $N_I = 1,000$ used in the previous simulation exercise we can see that any significant dependencies in the PAR samples will result in significant bias in the estimate of Inflows. Given the likelihood of a positive dependency between the PAR samples we can expect a significant over-estimate of Inflows for the class M2 estimator.

We also note here that the M5 class of estimator also depends on the M2 estimator class for adjusting a naive estimate of the 2 year population. If we undertake the same sensitivity analysis for M5, we will find the same level of bias for each value of θ. This is because we use the estimate of N_S from the M2 class of estimators to adjust the naive estimate accordingly. In essence, the M5 class of estimator is simply the naive M3 class of estimator recalibrated using the value of $\hat{N}_S$ obtained with the M2 class of estimators.

5.4.4 Learning points from the simulations

The key learning points obtained from the simulation exercises are as follows:

- The value of using these methods to estimate population flows that are relatively very small (relative size less than 5%), will very much depend on the scale (population size) and coverage (PAR coverage in the population).
- On looking at the simulation results in tables 5.3, 5.4 and 5.5, the DSE based methods seem to outperform the hypergeometric based methods in terms of precision across each of the methods. This difference also seems to be quite large, such

that, there is no benefit to combining DSE and hypergeometric based estimators. While the simulations did not examine the performance of the different estimators over different lists and population sizes, the increased precision of the DSE based estimates is explained by the high coverage rates in the PAR.

- the M2 and M5 classes of estimators seem to perform better than the M1 and M4 classes of estimators, respectively, in terms of accuracy (bias). However, a consideration of the sensitivity analysis results in table 5.7, where the sensitivity of method $M2_{(DSE)}$ to different dependency ratios between PAR_1 and PAR_2 is assessed, suggests a high degree of caution be taken with respect to the underlying assumptions.
- Because of the simplicity of the proposed estimator, $M1_{(DSE)}$, compared to the $M4_{(DSE)}$ estimator, given no other major differences, $M1_{(DSE)}$ will be preferred over $M4_{(DSE)}$.
- The results of the data simulation were only presented for one set up. The set up chosen is similar to that of young Irish males/females in a single year age group. While all population groups provide challenges with respect to estimating population flows, this is the largest group and has the least tangible information or evidence available with respect to population flows.

The descriptions of each of the estimators are contained in table 5.1.

5.5 Real World Application

To estimate the population flows, the same data sources are used to compile the population estimates as described in Zhang and Dunne (2018). The population estimates have already been presented in figures 3.14 and 3.15 and are compiled by nationality group, single year of age and sex before aggregating. Blocking in this manner, while allowing estimates to be compiled for natural groupings, also overcomes difficulties that may arise where different population groups have a different propensity to hold an Irish Drivers licence (this difference arises out of different rules for nationalities and sometimes for age groups).

The computation is complicated slightly through having to age groups from one year to the next. The computation can be quickly validated with spot checks using the identity $N_1 - N_O + N_I = N_2$.

The most problematic age groups in estimating flows are those between 20 and 44 where nearly all of the migration (inward and outward) is accounted for. When population flows are comprised mainly of births and deaths, the estimation of flows is not such a big issue, as they can be compiled more or less from administrative data sources.

Year	2012	2013	2014	2015	2016
Official estimate	130,500	132,100	134,900	142,300	147,700
$M1_{(DSE)}$	111,300	132,300	148,100	143,400	154,900
$M1_{(HYP)}$	120,600	134,000	138,800	173,000	176,900
$M2_{(DSE)}$	419,400	413,300	410,800	413,200	446,300
$M2_{(HYP)}$	382,400	360,200	337,300	424,500	428,800
$M3$	151,200	175,100	184,500	211,900	219,400
$M4_{(DSE)}$	117,500	135,500	149,200	152,400	154,500
$M4_{(HYP)}$	128,600	146,500	148,200	178,000	171,200
$M5_{(DSE)}$	354,000	351,100	347,500	365,200	339,200
$M5_{(HYP)}$	325,600	313,600	294,800	371,300	337,500

Table 5.8: Comparison of State level estimates of Inflows for different classes of estimator. State level estimates also included. Official estimates are with reference to the 12 months preceding mid april of the reference year. (Source for Official estimate: http://www.cso.ie Statbank Table ID PEA15, accessed 7th September 2018).

Estimates for each of the methods described in table 5.1 are made for single year of age, sex and broad nationality group (Ireland, UK, EU15 excluding UK and Ireland, EU25 excluding EU15, EU28 excluding EU25 and All nationalities excluding EU28). These estimates are then aggregated up to estimate broader categories. This approach is chosen over an approach where broader categories are calculated directly, to ensure coherence when estimating categories that are aggregations of single year age categories. Where estimates can be compiled with reference to either estimating the Stayers in the SPD for year 1 and also for year 2 a simple average of both estimates is used when comparing methods. Table 5.8 presents these results (Inflows) at State level, along with the official published figures.

The Official estimates in table 5.8 are with reference to the 12 month period prior to mid April in the reference year.

In considering the PECADO flow estimates in table 5.8 and comparing them to the Official estimates, we see that the M1 and M4 class of estimator appear plausible. We also see that the hypergeometric based estimates also generally provide for slightly higher estimates than the DSE based estimates. Estimates for M2 and M5 classes look to be significantly higher than those for M1 and M4. We note the sensitivity analysis results table 5.7, where a positive odds ratio will lead to an underestimate of Stayers (and hence an overestimate for Inflows), and suggest that this is a result of a positive dependence between PAR_1 and PAR_2. We will discount the M2 and M5 class of estimators for this reason as we have not identified any way to correct for this dependence.

We also discount the M3 class of estimator due to the methodological flaw giving rise to a biased estimator. However, we note in the simulation exercises in section 5.4.2 that M3 class gave significantly lower estimates for Inflows across all the classes, whereas, in table 5.8, the estimates for M3 class are higher than those for M1 and M4 classes and below those for M2 and M5 classes. This may require further investigation.

This leaves one candidate method, M1, with two variants, the DSE based variant and the hypergeometric based variant. We noted from our simulations earlier that the DSE based method has slightly better precision. For this reason, and the property that the variance and bias tend to 0 as the SPD coverage goes to 100%, we dismiss the hypergeometric based method in favour of the DSE based method.

We now consider the distribution of the gross population flows in figure 5.8 by gender and single year of age in the 20 to 50 age group for reference year 2015. We consider both the year 1 and year 2 based estimates. Both the year 1 (red) and year 2 (blue) based estimates appear to be erratic from year to year with no discernible difference between either set of estimates across age and sex. In reality, we would not expect the true values for population flows to be so erratic over single year of age. The estimates, when considered over broader age categories are plausible with far higher flows estimated in the younger age categories than in the older age categories.

Figure 5.9 considers the estimates of gross population flows over time (2011 - 2016) and for broad age categories by gender. Again red represents the year 1 based estimates and blue represents the year 2 based estimates. The first row of plots, with reference to the age category 25 - 44 years, also includes the Official estimates published by CSO in black. Continuous lines represent Inflows while dashed lines represent Outflows. Plotting Inflows and Outflows on the same graph facilitates a consideration of net flows. In considering the proposed sets of estimates with the official produced statistics, the proposed estimates are of a similar if slightly higher magnitude. This indicates that the proposed estimates are plausible over time. When comparing the year 1 based estimates (red) and the year 2 based estimates (blue) we see that the two sets of estimates are broadly similar, with the year 1 based estimate tending to be a slightly lower more often than not. For females, graphs in right hand column of figure 5.9, there is a noticeable peak in Outflows for reference year 2013 and a corresponding peak in Inflows for reference year 2014 occurring in age categories 35 - 39 years and 40 - 44 years. These peaks carry into the 25 - 44 years age category. We investigate this feature a little further in figure 5.10 and notice exceptional peaks in Outflows for reference year 2013 in the categories United Kingdom and All nationalities excluding EU 28 with a corresponding peak in Inflows for reference year 2014 in the Ireland category. There is no apparent corroborating evidence or thinking to believe these peaks correspond to real world events related to Outflows and Inflows and surmise that the peaks probably come about as a result of new administrative routines that provided an opportunity for persons to update information about themselves. In particular, a new Public Services Card was launched at around this time for those engaging with certain public administration systems (driver licence, social welfare). This points to a weakness in the method if the blocking variables are not stable over the two reference periods used to estimate flows. The primary cause of the instability will originate from changes in the composition of the population estimates from one year to the next due to a change in nationality attribute from one year to the next. If we remove or adjust these peaks downwards then, referring back to figure 5.9 and considering the females 25 - 44 years age category for all nationalities, the estimates produced by the proposed methodologies will correspond much better to the Official estimates - the adjustment to the peaks being somewhere in the region of 10,000 persons. However, it is difficult to formally reduce these peaks as then the coherence of flows with stock estimates in the system of population estimates would be incoherent at broad nationality grouping. This points to a weakness in the system if blocking variables (nationality group) are not consistent over time. Practical strategies to overcome this weakness could include removing the nationality group as a blocking variable when compiling estimates or acknowledging the change of nationality grouping as being included in the estimates of Inflows and Outflows. While not explored here, the tests described in section 2.4 can be used to undertake this evaluation.

We use a data simulation or bootstrap exercise to explore the level of variation in the proposed flow estimates. The data simulation exercise is set up by first creating the populations and SPDs as estimated using the PAR and DLD for years 1 and 2. We use the estimates of Stayers and population size estimates at year 1 and year 2 as the true population values. We then simulate the PAR datasets as described by the estimates knowing the size of PAR_1 and PAR_2 and the size of the overlap $PAR_1 \cap PAR_2$, and we use the estimate of the Stayers in PAR_1 and PAR_2 as the basis for specifying how many persons from the population of Stayers and non-Stayers should be included in each of PAR_1 and PAR_2. We treat PAR_1 and PAR_2 as fixed and then draw repeated samples to represent DLD_1 and DLD_2 from populations 1 and 2 respectively and calculate estimates each time a draw is made. Each person in the population is assigned to one of two waves, such that, the samples DLD can be drawn from different waves to ensure no overlap. The variation in the estimates over each draw will then provide insights into the variation in the estimates themselves. To properly replicate the real world application and get a good indication of variation in the estimates, 1,000 estimates were compiled for each combination of single year of age, sex and nationality grouping for reference year 2015. We then aggregated over categories to obtain estimates of variation for any summary category. Table 5.9 presents the co-efficients of variation alongside the inflow estimates by age group and gender for reference year 2015.

From table 5.9, we can see that for the 5 year age groups the variation is somewhat acceptable varying from a CV of approximately 5% when estimating larger quantities in the 25 to 29 age group to a CV of approximately 15% when estimating smaller quantities in the 40 to 44 age group. When looking at an estimate for the entire age group 25 to 44

Age Group	Estimate	Variance	SE	CV (%)
Males				
25 - 29	10,090	267,940	520	5.1
30 - 34	7,780	299,960	550	7.0
35 - 39	4,280	197,500	440	10.4
40 - 44	3,860	311,350	560	14.5
25 - 44	26,010	1,076,740	1,040	4.0
Females				
25 - 29	8,950	207,260	450	5.1
30 - 34	5,020	95,990	310	6.2
35 - 39	3,070	70,440	260	8.6
40 - 44	1,830	84,670	290	15.9
25 - 44	18,870	458,370	680	3.6

Table 5.9: Results of data simulation exercise to explore variation in the estimates of population inflows for reference year 2015.

	Bootstrap	Estimated	
Age Group	Variance (v_b)	Variance (v_e)	$\frac{\sqrt{v_e}}{\sqrt{v_b}}$
Males			
25 - 29	371,671	621,613	1.29
30 - 34	303,766	383,810	1.12
35 - 39	234,065	364,740	1.25
40 - 44	207,377	299,371	1.20
25 - 44	1,116,879	1,669,534	1.22
Females			
25 - 29	281,469	402,863	1.20
30 - 34	134,943	189,485	1.18
35 - 39	89,221	125,005	1.18
40 - 44	71,672	115,346	1.27
25 - 44	577,305	832,699	1.20

Table 5.10: Comparison of Bootstrap Variance (v_b) versus Estimated Variance (v_e) from equation 5.5 for N_S. Reference year 2015.

the estimate becomes more reliable with a CV in the region of 4%. As we would expect to see, the smaller the quantity we are trying to estimate the less stable the estimate will tend to be.

We can also consider the empirical variance from the simulations and the estimated variance for N_S in table 5.10 and again see that the estimated variance has a considerable level of conservatism involved when compared with the bootstrap variance. The estimated variance, as per equation , is conservative in its derivation. In addition, the estimated variance does not account for situations where the estimates are truncated when they lie outside the bounds of possibility. Using equation 5.5 to estimate the variance, as opposed to the empirical bootstrap approach, will result in overestimating the standard error of the estimates by between 12% and 29% for the sub-population groups presented in table 5.10. Therefore, we conclude that if it is feasible to compile variance estimates using boostrap methods as done here, in the absence of a more precise estimator, it is preferable.

5.6 Concluding remarks

In Ireland, population estimates depend on the previous year's estimate and the ability of the demographic component method to bring those estimates forward another year. Ireland does not have a Central Population Register(CPR) it can rely on for these estimates. Reliable estimates of births and deaths are not so difficult when there are efficient administration systems to capture this information. However, it is very difficult to obtain reliable estimates of migration flows and, in a small open economy like Ireland, these flows can have a sizeable impact on population estimates. Bringing forward population estimates year by year using the demographic component method introduces increasingly more potential error for each additional year it is carried forward. This is one of the reasons Ireland conducts a Census every 5 years instead of every 10 years, at a significant added cost.

To overcome the reliance on the demographic component method for population estimates, a new approach to annual population estimates, based on creating an incomplete statistical register of persons and using a DSE approach to adjust for undercoverage, has been previously suggested and is currently being explored by the CSO. If this method can provide reliable annual population estimates, then, by definition, it will also provide reliable estimates of net population flows. However, there is still significant statistical value in understanding gross population flows to obtain insights into migration. Without complete registers or perfect border checks for entries and exits, there are no existing methods known to the author that are being used to compile gross population flows from administrative data only. This chapter has developed and proposed a method to estimate gross population flows with incomplete SPDs using administrative data only. The chapter has also developed and presented a number of alternative methods and evaluated them against the proposed method showing it to be the most viable method of those evaluated.

Finally, it should be noted that these methods may have practical applications beyond estimating population flows with incomplete registers. This list approach, for one, may have applications in situations as described in Chao et al. (2008). The methods described here may prove to be superior to the Chao method where the bias introduced by these methods can be dealt with or is not considered a significant issue. The methods may also have practical applicability where the assumptions or conditions associated with the Chao method may not hold.

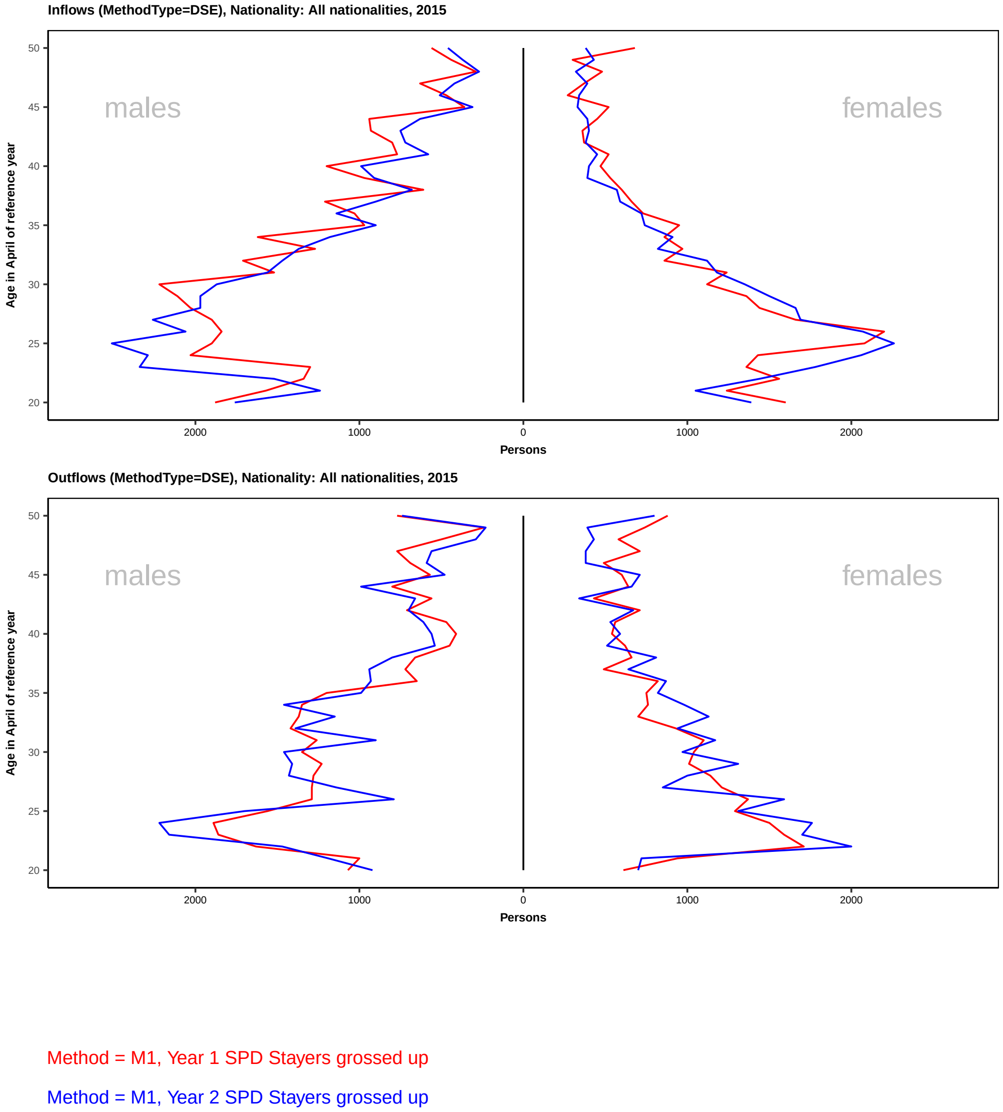

Figure 5.8: DSE based (M1) estimates of gross population flows by gender and single year of age (20 - 50 years), reference year 2015. Red represents estimates based on grossing up an estimate of Stayers in the SPD in Year 1. Blue represents estimates based on grossing up an estimate of Stayers in the SPD in year 2. For Outflows, years 1 and 2 represent 2015 and 2016. For Inflows, years 1 and 2 represent 2014 and 2015.

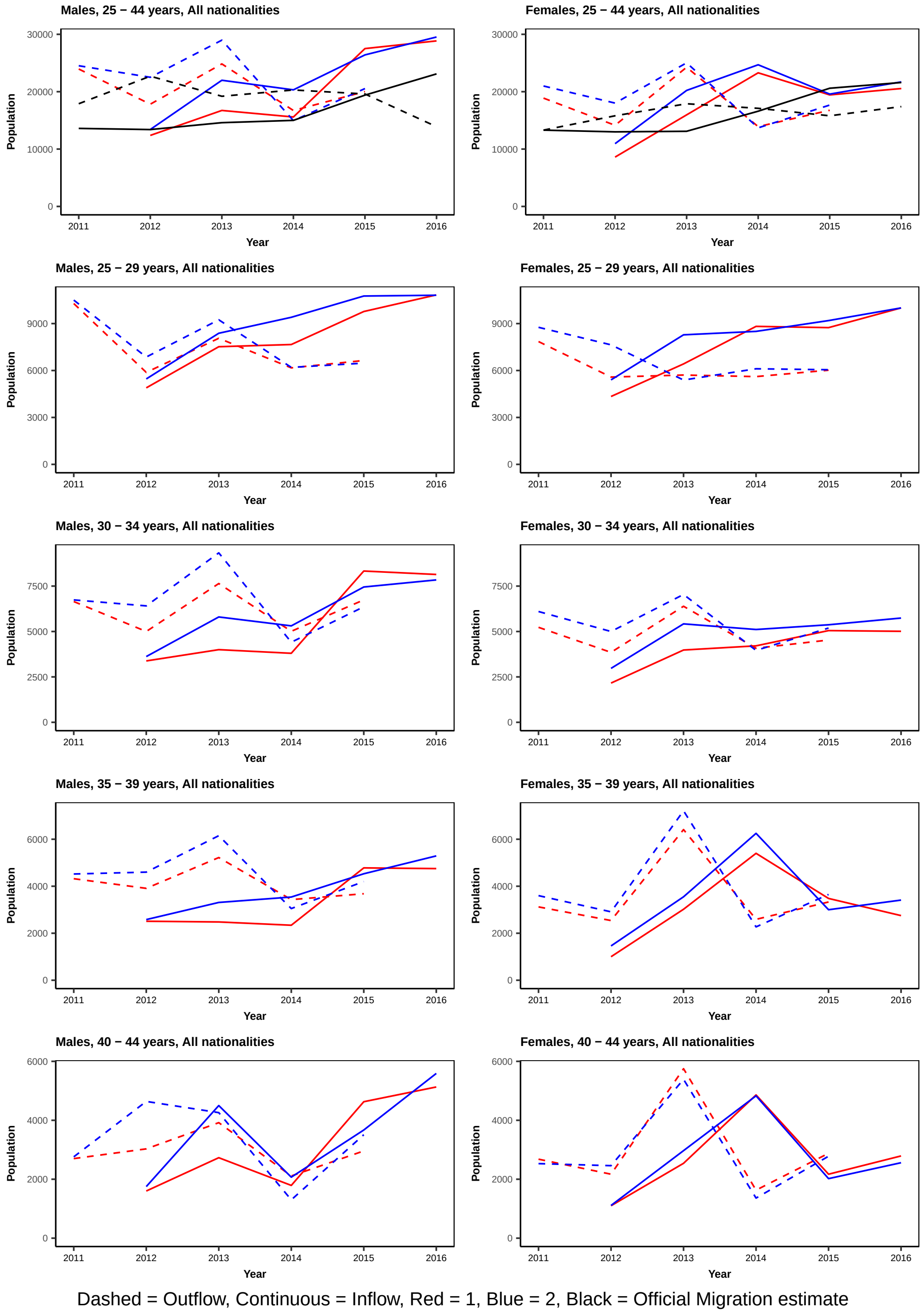

Figure 5.9: DSE based (M1) estimates of gross population flows by age group, gender and reference year (2011 - 2016). Red represents estimates based on grossing up an estimate of Stayers in the SPD in Year 1. Blue represents estimates based on grossing up an estimate of Stayers in the SPD in year 2.

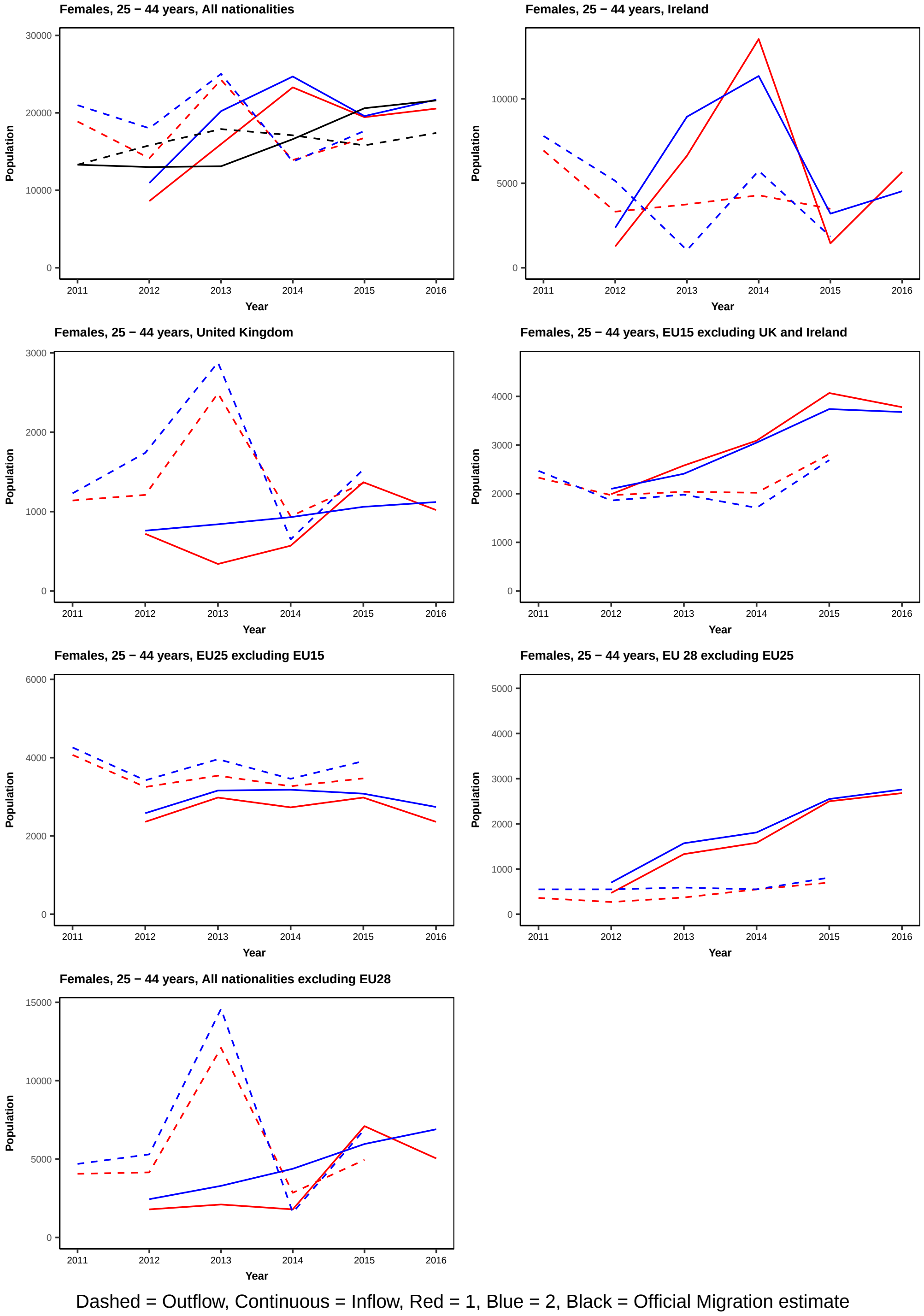

Figure 5.10: DSE based (M1) estimates of gross population flows for females by age group, nationality group and reference year (2011 - 2016). Red represent estimates based on grossing up an estimate of Stayers in the SPD in Year 1. Blue represents estimates based on grossing up an estimate of Stayers in the SPD in year 2.

Chapter 6

Conclusions

6.1 Overview

The Irish PECADO project has been undertaken as a follow up to the initial steps outlined in the *revolutionary approach* described in Dunne (2015). While Dunne (2015) describes a Census opportunity, it can be argued that since that paper the Census environment has changed significantly. For many countries, the traditional Census model is no longer considered fit for purpose, given the cost and regularity at which a Census is undertaken. There is also an increased challenge with respect to response for the traditional Census model along with an emerging requirement for Census-like population estimates broken down by detailed geography on an annual basis.

Chapter 3, underpinned by methodology presented in chapter 2, proposes a system of annual Population Estimates Compiled from Administrative Data Only (PECADO). The PECADO system, as presented, provides for population estimates broken down by single year of age and gender. While blocking by nationality group is undertaken, we have refrained from proposing that the system publishes estimates broken down by this variable for coherence and conceptual reasons. The nationality attribute, while generally complete on administrative data sources, will have been collected for administrative purposes at a point in time (usually at a first registration) and probably not updated on any regular basis. Nationality as collected by the Census (or a household survey) can differ from that contained in administrative sources due to timing and mode or context effects. The following scenario illustrates this difference; a person moves to Ireland from the United Kingdom and applies for an identification number, whereupon their nationality is recorded as United Kingdom (maybe due to evidence of previous address or passport), that person subsequently marries or integrates into Irish society and feels Irish and, as such, when completing a Census form or responding to an interviewer will declare themselves to be Irish. This is a common challenge when using attributes from administrative data that may be subject to timing, definition or mode effects. Age and gender will not be impacted in the same manner.

We propose the PECADO system of population estimates as a strong foundation for a new system of population estimates, given high quality linkage between data sources and the considerable efforts taken to identify erroneous records. We see from table 3.3 that the PECADO population estimates differ from the Census (usual resident) counts by 5.5% in 2011 and 6.3% in 2016. We note slight differences in the concepts; the Census (usual resident) count relates to persons usually resident and present on Census night whereas the underlying PECADO population concept relates to persons usually resident at any particular time in the year. Therefore, by definition, the PECADO estimate will be greater than the Census count as it will also include

- persons who would normally be considered usually resident and who emigrated at some point between the start of the reference year and Census night (not present on Census night)
- persons who would normally be considered usually resident and immigrated at some point between Census night and the end of the reference year (not present on Census night)
- persons considered usually resident on Census night but for some reason (travelling abroad) were not present on Census night

We know from official migration estimates, see table 1.1, that migration alone will not fully reconcile the differences between the population concepts and suspect that under-coverage may be an emerging problem for the traditional Census model in Ireland. CSO will incorporate a UCS into the Census in 2021.

Chapter 4, again underpinned by methodology presented in chapter 2, explores the possibility of using the same DSE methods to adjust the Census usual resident count for undercoverage. The added innovation in this approach is the use of a second list compiled from administrative data rather than a list compiled using a field survey. This approach has the possibility of reducing associated UCS time and costs significantly when compared to the traditional UCS model involving a field collection. Statistics New Zealand (SNZ) recently incorporated this idea of using an administrative list into their 2018 Census to compile reference population estimates. SNZ will subsequently use the Post Enumeration Survey (PES) as part of a quality assurance exercise at a later date (Bycroft et al., 2018). The PES typically includes a UCS component. As detailed in section 4.4, we determine an undercount of 5.2% when taking the population concept as those usually resident in the month of April. When we compare these Census figures adjusted for undercoverage with the PECADO estimates in figure 4.2 we see the primary differences in the estimates occurring in the young adult part of the population, that part of the population more likely be part of migration flows. The analysis suggests undercoverage in the Census 2016 population as an explanation for the difference between the PECADO estimates and the Census counts, however, there is also a MAR assumption applying to 733,800 persons not on the *trimmed* Census dataset or nearly 15% of the final population estimate. There is a weakness in the traditional Census model in Ireland which relies on the field force to fully enumerate the population, allowing no adjustment for undercount or no reassurance that undercount has been eliminated. The difference between the PECADO estimate and the Census estimate is greater in 2016 than in 2011, table 3.2, indicating that if the difference is due to undercoverage, then undercoverage has increased over time.

A key Census requirement is the ability to be able to provide population estimates at a detailed geographical level. The PECADO project to date has not addressed this. In exploring the possibility of conducting an undercoverage survey at a detailed geographical level with these methods, it was found that *address mislocation* was a potential source of error that would need to be addressed. *Address mislocation* occurs where a person is assigned to different addresses in each of the two lists. A similar *address mislocation* challenge will also need to be overcome to disaggregate the PECADO population estimates derived in chapter 3.

Reliable population estimates for two periods will, by definition, give reliable estimates of net flows between those periods. The net flows are simply obtained by differencing. However, there is also a significant interest in gross flows, inflows comprising of immigration and births and outflows comprising of emigration and deaths. In chapter 5 we propose and evaluate an extension of the PECADO system to estimate gross population flows. The PECADO gross flow estimates are plausible when considered alongside official estimates and are by design coherent with the PECADO population estimates produced using the same data sources.

The PECADO system provides for a coherent system of stocks and flows. The primary advantage of this system over the existing official system of estimates is that it does not require post censal estimates to be revised after each Census (which in the Irish case can be significant). Such revisions have an impact on a large volume of statistical products where population estimates are used to compile weights, estimate rates or provide other context to published statistics. A further weakness in the existing system of population estimates is the inability to provide robust estimates of population flows that are in turn used to compile the subsequent period's population estimate. This last weakness is the primary reason why countries such as Ireland conduct a Census every 5 years.

6.2 Considerations and Insights

6.2.1 Population concepts

Historically, in the traditional Census setting, the preferred population concept has been the 'de facto' population concept which refers to the number of people that were present on a given night. Register-based countries, discussed in section 1.2.1.2, would typically use the concept of a legally resident population on a given night. For international comparisons, the preferred concept is typically the usually resident population on a given date, with the date being allowed to vary according to country specific preferences or requirements. One of the key drivers behind the population concept is to prevent double counting or omissions across regions when conducting a Census. In the traditional context this is done by counting everybody in place on a given night, hence a de facto type definition. In considering the traditional register-based countries, where persons are registered to an address, the legally resident concept (whether related to a point in time or period) is more suited, as a person has stated this is where they can be contacted, double counting is prevented through the use of official identification numbers. The recommended concept for international comparison purposes is typically based on a usual residence population concept, relating to where a person has been resident and/or has an intention to continue to be resident for a period of 12 months. Lanzieri (2013) provides a more in depth discussion on population concepts and considers a new population concept - *the annual resident population.*

The annual resident population concept is a concept that closely underpins the population estimates in the PECADO project. The PECADO project uses a SoL based on a significant activity that equates to someone being resident for a significant period of time in the reference year. The choice of population concept was chosen based on which concept would be easier to implement and understand. If there is a requirement to compile population estimates by a different population concept, it is far easier to identify the components of difference, estimate them and adjust the *reference* population estimates accordingly. For example, if the PECADO project is required to produce an estimate of the usually resident population at a point in time then this can be done by subtracting from the PECADO estimate an estimate for outflows prior to the point in time and an estimate for inflows subsequent to the point in time. If the point in time is the first of January then the adjustment is simply the estimate of inflows for the reference year. This approach also makes it much easier to ensure a more coherent approach when compiling population estimates underpinned by different concepts.

6.2.2 Timeliness

Timeliness is a critical consideration in the production of Official Statistics.

In considering the underlying data sources used in the PECADO system, see table 3.1, it is expected that robust population estimates for a reference year could easily be compiled by the middle of the following year. Only one data source, self-employed income tax returns, used in the PECADO system may not be available due to reporting deadlines. The availability of data sources also depends on CSO maintaining strong positive relationships with data suppliers. In terms of the production of such Official Statistics, 6 months after the end of the reference period should generally be considered acceptable.

This timeline may not be considered good enough for some statistical systems where population estimates play an integral role. For example, weights are required for household surveys such as the Labour Force Survey (LFS) and the Survey on Income and Living Conditions (SILC). However, using the previous year's PECADO estimate and rolling forward with estimated counts of births, deaths and net migration flows, reasonable population *nowcasts* can be obtained for use in other statistical systems. If required, the statistics created using *nowcasts* can be revised the following year when the new population estimates become available. It would be desirable, and may even be probable, that the difference between the *nowcasts* and the estimates when produced are small enough to render any potential revision negligible and irrelevant. This approach would be preferable to users as it would render redundant the large volume of revisions over a large number of years across a broad group of Official Statistics involved with the traditional approach. The traditional approach involves revising postcensal population estimates with intercensal population estimates after each Census. With the 5-yearly Census in Ireland, this involves revising official statistics over a period of 6 to 7 years retrospectively.

The proposed PECADO system holds the promise of no significant post censal revisions to population estimates and the accompanying domino effect of revisions to all Official Statistics that incorporate population estimates in their production processes.

Under the current demographic cohort approach to producing population estimates, the compilation of migration estimates is critical in the compilation of population estimates. However, in the proposed PECADO system, the compilation of populations flows is undertaken as a secondary exercise and depend first on having robust population estimates. Once the PECADO system has provided population estimates for Years 1 and 2, then estimates of population outflows for Year 1 and population inflows for Year 2 can be compiled using the same data sources. The estimate of population outflows will be 1 year behind the estimate of population inflows.

Undertaking a UCS using administrative data will have significant benefits in terms of timeliness if the administrative data source is available in near real time and the linkage can be done as Census responses are recorded. The administrative data source used in chapter 4 is available the week after payments are made. In fact, it should be possible to compile population estimates adjusted for undercoverage at the end of each day or week and watch the population estimates and Census counts converge as the response rate increases. Incorporating such a feature into Census processing can only contribute to the quality assurance of the Census processing systems. There still remains the challenge of disaggregating any identified undercoverage by detailed geography.

6.2.3 Methodology

The methodology considerations in chapter 2 allow for a far broader application of DSE methods in the production of Official Statistics and were driven by 3 key insights.

The first key insight is to frame the problem such that, it is easy to understand and, similarly, the methodology to be employed is also easily explained. A first natural approach to population estimation with so many underlying data sources would be to consider each of the data sources in their own right with their own characteristics. However, a key innovation was to focus on a Signs of Life (SoL) approach and integrate the underlying data sources into a SPD with only one record per person where SoL exist. This meant the SPD by design had only undercoverage, linkage error is eliminated through the use of high quality linkage keys and erroneous records by definition should be removed if the SPD is restricted to records with high quality SoL. A second innovation was then to consider using a suitable administrative list as list B in a DSE setup.

A key principle underpinning the data sources in the PECADO project was to only use records where there was high or absolute confidence that the record represented the person it was suppose to. A rule of thumb, *if in doubt leave it out*, ensured a simplification and streamlining of the challenge in producing population estimates from administrative data sources only. A similar idea is implicit in the Spanish 2010 Census, discussed in section 1.2.1.3.

The second key insight is that the traditional DSE assumptions, as presented in Wolter (1986), can be relaxed to 3 basic assumptions labeled in section 2.2.2 as *No erroneous records*, *Matching assumption (no linkage error)* and *Homogeneous capture with respect to list B*, with a fourth assumption of *Independent Capture (list B)* added to derive an expression for the variance of the DSE estimator. This approach, with particular emphasis on the homogeneous capture assumption applying to list B, allows a much broader application of DSE methods, particularly in situations where one or both lists may be compiled from administrative sources. In fact, when heterogeneity in capture rates is explored in section 2.4, we see from equation 2.16, that under certain conditions the assumption, *Homogeneous capture with respect to list B*, can be relaxed even further without introducing bias. If list B is drawn from 2 sub-populations with different capture rates, then if the coverage rates of list A with respect to the two sub-populations are equal, no bias will be introduced. This result provides additional reassurance in the use of the Driver Licence Dataset as a list B in estimating the population size; we expect no difference in the behaviour of drivers and non-drivers when interacting with Public Administration Systems. This result further broadens the applicability of DSE methods as demonstrated in chapters 4 and 5. Zhang (2019) provides further consideration of the assumptions underpinning Dual System Population Size Estimator.

The third key insight is the extension of DSE methods to evaluate list A for erroneous records. Trimmed DSE, presented in section 2.3, is a valuable addition to the PECADO toolkit and DSE methods generally, as it provides a way to evaluate suspect parts of list A for the presence of erroneous records provided the remaining assumptions hold (i.e., no linkage error, no erroneous records in list B, and homogeneous capture in list B). This opens up the possibility of eliminating costly surveys purposely designed to deal with overcoverage in systems of population estimates, thus reducing costs and simplifying systems and explanations for compilers and users.

6.3 Next steps for PECADO project

In terms of producing *Census-like* population estimates, the PECADO project needs to address three further areas, namely

Geography The PECADO population estimates are currently produced at State level. There is a requirement for them to be compiled at a detailed geographical level. This is the highest priority.

Attributes The PECADO population estimates do not provide any breakdown beyond gender and single year of age. There is a requirement to provide a break down of the population by a number of other attribute classifications: citizenship, marital status, ILO status, to name three.

Household Composition The PECADO population estimates do not provide any analysis of housing occupancy or household composition. There is a requirement to provide this analysis.

There are challenges with each of the above areas, primarily to do with quality, conceptual definitions and coherence, when compiling this information directly from different administrative data sources. For example, addresses or marital status may be recorded differently for the same person on different sources for any number of reasons. New methods need to be developed or existing methods need to be adapted to overcome these challenges.

There is methodological research into extending DSE methods to address issues of incomplete covariate (attribute) information or misalignment between the same classifications in two sources, provided the classification on one source can be considered the truth (or target classification concept) (van der Heijden et al., 2018). Combinations of attributes (nationality by marital status) may also be accommodated in this approach. This research may also have applicability to the *address mislocation* problem and household composition challenge.

There are also ongoing discussions within the PECADO project about how a *calibrated* DSE could be developed and employed to adjust for *address mislocation.*

Previous research work on a *unit error theory* (Zhang, 2011) may also form the basis of a solution to the household composition challenge, whereby the approach would be to form household units from the available person units in the SPD using relationship or other linking information in administrative data sources.

The above approaches consider the three challenges in isolation, consideration needs also to be given to considering these challenges together where there is a requirement to do so. For example, consider a requirement for population estimates by household composition and geography.

Depending on how the challenge of disaggregating population estimates by geography is solved, it may also be possible to incorporate the approach taken in chapter 5 to provide annual estimates of internal migration.

In summary, the PECADO project has made a lot of progress in compiling a reference population estimate at State level (for Ireland) by gender and single year of age. However, this work needs now to be built on to meet the additional requirements associated with Census-like population estimates broken down by detailed geography on an annual basis.

References

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

Argüeso, A. and Vega, J. L. (2014). A Population Census Based on Registers and a 10 % Survey Methodological Challenges and Conclusions. *Statistical Journal of the IAOS*, 30:35–39.

Baffour, B., Brown, J. J., and Smith, P. W. F. (2013). An Investigation of Triple System Estimators in Censuses. *Statistical Journal of the IAOS*, 29(1):53–68.

Bechtold, S. (2016). The 2011 Census Model in Germany. Last accessed on 17/06/2020 at https://www.comparativepopulationstudies.de/index.php/CPoS/article/view/217/227.

Beltadze, D. (2020). Developing Methodology for the Register-based Census in Estonia. *Statistical Journal of the IAOS*, 36:159–164.

Bengtsson, T. and Rönning, S. Å. (2016). Overcoverage in the Total Population Register. In *Nordiskt Statistikermöte - Statistics in a changing world. Towards 2020 and beyond*, page 12, Stockholm. Statistics Sweden.

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

Blum, O. and Feinstein, Y. (2017). Estimation of the Total Population in the 2020 Integrated Census in Israel. In *UNECE Group of Experts on Population and Housing Censuses*. UNECE. Last accessed on 12th May 2020 at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2015/mtg1 /CES_GE.41_2015_4-Israel_rev.pdf.

Bycroft, C., Connolly, K., and Quinn, A. (2018). Transcript : 2018 Census Technical Seminar. Last accessed on 14th April 2020 at https://www.stats.govt.nz/methods/2018-census-how-we-combined-administrative-data-and-census-forms-data-to-create-the-census-dataset.

Central Bureau of Statistics of Israel (2015). The First Round of the Rolling Integrated Census in Israel Methodology, Results and Flaws. In *UNECE Group of Experts on Population and Housing Censuses*. UNECE. Last accessed on 17/06/2020 at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2015/mtg1/CES_GE.41_2015_4-Israel_rev.pdf.

Chao, A. (2015). Capture-Recapture for Human Populations. *Wiley StatsRef: Statistics Reference Online*, pages 1–16. Last accessed on 17/06/2020 at http://doi.wiley.com/10.1002/9781118445112.stat04855.pub2.

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.

Chapman, D. G. (1951). Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses. *University of California Publications in Statistics*, 1:131–160.

CSO (2003). Statistical Potential of Administrative Records An Examination of Data Holdings in Six Government Departments Working Report Central Statistics Office. Technical report, Central Statistics Office, Ireland. last accessed on 17/06/2020 at http://www.cso.ie/en/media/csoie/releasespublications/documents/otherreleases/spar.pdf.

CSO (2006). Statistical Potential of Business and Environment Enterprise Data Holdings in Selected Government Departments Working Report Central Statistics Office. Technical report, Central Statistics Office, Ireland. Lat accessed on 17/06/2020 at http://www.cso.ie/en/media/csoie/releasespublications/documents/otherreleases/spar_bes.pdf.

CSO (2009). Statistical Potential of Administrative Records An Examination of Data Holdings in the Office of the Revenue Commissioners Working Report. Technical report, Central Statistics Office, Ireland. Last accessed on 17/06/2020 at http://www.cso.ie/en/media/csoie/releasespublications/documents/corporatepublications/CSO_Revenue_SPAR.pdf.

CSO (2017). *Census 2016 Summary Results - Part 1*. Central Statistics Office, Ireland. Last accessed on 17/06/2020 at https://www.cso.ie/en/media/csoie/newsevents/documents/census2016summaryresultspart1/Census2016SummaryPart1.pdf.

DES (2013a). Early Leavers What Next? Report on Early Leavers from Post-Primary schools. Technical report, Department of Education and Skills. Last accessed on 17/06/2020 at https://www.education.ie/en/Publications/Statistics/Early-Leavers-What-Next-.pdf.

DES (2013b). School Completers What Next? Report on School Completers from Post-Primary Schools. Technical report, Department of Education and Skills. Last accessed on 17/06/2020 at http://www.education.ie/en/Publications/Statistics/School-Completers-What-Next-.pdf.

DPER (2011). Public Service Reform. Last accessed (web page) on 05/05/17 at http://www.per.gov.ie/en/public-service-reform/.

Dunne, J. (2011). Exploiting Administrative Data to Investigate where those Leaving Jobs get Re-employed. In *58th World Statistical Congress*, pages 1888–1897. International Statistical Institute. Last accessed o 17/06/2020 at http://2011.isiproceedings.org/papers/450470.pdf.

Dunne, J. (2015). The Irish Statistical System and the Emerging Census Opportunity. *Statistical Journal of the IAOS*, 31(3):391–400.

Dunne, J. and Graham, P. (2019). New Population Estimation Methods : New Zealand and Ireland. In *ISI World Statistics Congress 2019*, number August, Kuala Lumpur. Last accessed on 17/06/2020 at http://isi2019.org/proceeding/2.STS/STS%20VOL%203/index.html#p=323.

Durr, J.-M. (2005). The French new rolling Census. *Statistical Journal of the United Nations ECE*, 22:3–12.

Eichenberger, P., Potterat, J., and Hulliger, B. (2010). Describing the Anticipated Accuracy of the Swiss Population Survey. Technical report, Swiss Federal Statistical Office. Last accessed on 17/06/2020 at https://www.bfs.admin.ch/bfsstatic/dam/assets/347682/master.

EUROSTAT (2003). *Demographic Statistics: Definitions and Methods of Collection in 31 European Countries.* European Communities. Last accessed on 17/06/2020 at https://ec.europa.eu/eurostat/ramon/statmanuals/files/KS-CC-03-005-EN.pdf.

EUROSTAT (2015). *Demographic Statistics: A Review of Definitions and Methods of Collection in 44 European Countries.* Eurostat. Last access on 17/06/2020 at https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-15-002.

FSO (2015). New Census System Quality Survey. Technical Report January, FSO. Last accessed on 17/06/2020 at https://www.bfs.admin.ch/bfs/en/home/basics/census/quality-survey-national-census-system%20.assetdetail.350200.html (Federal Statistics Office).

Gallo, G., Chieppa, A., Tomeo, V., and Falorsi, S. (2016). The Integration of Administrative Data Sources in Italy to Increase Population Census Data Availability. In *UNECE Group of Experts on Population and Housing Censuses*, number September, pages 1–15. UNECE. Last accessed on 17/06/2020 at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2016/mtg1/CES_GE.41_2016_3E_Italy_rev_2.pdf.

Gerritse, S. C., Bakker, B. F. M., de Wolf, P. P., and van der Heijden, P. G. (2016). Under Coverage of the Population Register in the Netherlands , 2010. *CBS Discussion Paper 2016 — 02*, (February):1–31.

Gerritse, S. C., van der Heijden, P. G. M., and Bakker, B. F. M. (2015). Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models. *Journal of Official Statistics*, 31(3):357–379.

Goodman, L. A. (1960). On the Exact Variance of Products. *Journal of the American Statistical Association*, 55(292):708–713.

Hayes, J. and Dunne, J. (2012). Realising the Statistical Potential of Administrative Data. In *General Conference of European Statisticians, Seminar on New Frontiers for Statistical Data Collection*. UNECE. Last accessed on 17/06/2020 at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP16.pdf.

INE Spain (2014). Population Figures Methodology. Technical Report July. Last accessed on 17/06/2020 at https://www.ine.es/en/metodologia/t20/t2030321_en.pdf.

INE Spain (2018). Migration Statistics Methodology. Technical Report February. Last accessed on 17/06/2020 at http://www.ine.es/en/metodologia/t20/t2030277_en.pdf.

Jensen, E. (2013). A Review of Methods for Estimating Emigration. (US Census Bureau) Last accessed on 17/06/2020 at https://www.census.gov/content/dam/Census/library/working-papers/2013/demo/jensen-01.pdf.

Kamen, C. S. (2005). The 2008 Israel Integrated Census of Population and Housing. *Statistical Journal of the United Nations ECE*, 22:39–57.

Kish, L. (1995). *Survey Sampling*. Wiley-Interscience, revised edition.

Kraus, R. (2013). Statistical Déjà Vu: The National Data Center Proposal of 1965 and Its Descendants. *Journal of Privacy and Confidentiality*, 5(1).

Lange, A. (2014). The population and housing Census in a register based statistical system. *Statistical Journal of the IAOS*, 30(1):41–45.

Lanzieri, G. (2013). On a New Population Definition for Statistical Purposes Note. In *CES Group of Experts on Population and Housing Censuses*. UNECE. Last accessed on 18/06/2020 at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2013/census_meeting/Eurostat_introductory_paper_on_new_population_definition.pdf.

Linehan, T. P. (1992). History and Development of Irish Population Censuses. *Journal of the Statistical and Social Inquiry Society of Ireland*, XXVI(IV):91–132. Last accessed on 18/06/2020 at http://www.tara.tcd.ie/xmlui/bitstream/handle/2262/2762/jssisiVolXXVI91_132.pdf.

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

Maasing, E. and Tiit, E.-M. (2019). An Index Based Approach to Determine Estonian Population by using Administrative Sources. In *Population Census, 2020 Round and Post 2020: From Traditions to Modernism*, pages 1–11, Bucharest. Last accessed on 12th May 2020 at http://census.statisticsevents.ro/wp-content/uploads/2019/04/07_Estonia-Ethel-Massing.pdf.

Macfeely, S. and Dunne, J. (2014). Joining Up Public Service Information: The Rationale for a National Data Infrastructure. *Administration*, 61(4):93–107. Last accessed on 18/06/2020 at https://www.ucc.ie/en/media/academic/centreforpolicystudies/CPSWP13-010-MacFeelyS,DunneJ,(2013)JoiningUpPublicServiceInformation-TheRationaleforaNationalDataInfrastructure.pdf.

McNally, J. and Bycroft, C. (2015). Quality Standards for Population Statistics : Accuracy Requirements for Future Census Models. Technical report, Statistics New Zealand. Last access on 18/06/2020 at http://archive.stats.govt.nz/~/media/Statistics/surveys-and-methods/methods/research-papers/topss/quality-stds-pop-stats.pdf.

NISRA (2015). Northern Ireland Census 2011 Quality Assurance Report. Technical Report March, The Northern Ireland Statistics and Research Agency. Last accessed on 18/06/2020 at https://www.nisra.gov.uk/sites/nisra.gov.uk/files/publications/2011-census-quality-assurance-report.pdf.

Nordbotten, S. (2010). The statistical archive system 1960-2010: A summary. *Nordisk Statistikermøde i København*, 11. Last accessed on 18/06/2020 at https://www.researchgate.net/publication/283643949_The_statistical_archive_system_1960-_2010.

Nordholt, E. S. (2005). The Dutch Virtual Census 2001 : A New Approach by Combining Different Sources. *Statistical Journal of the United Nations*, 22:25–37.

Nordholt, E. S. (2017). Draft UNECE Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses. In *CES group of Experts on Population and Housing Censuses*, number October. UNECE. Last accessed on 18/06/2020 at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/GE_41_2017_20_ENG.pdf.

Nordholt, E. S., Van Zeijl, J., and Hoeksma, L. (2014). *Dutch Census 2011: Analysis and methodology.* Statistics Netherlands. Last accessed on 18/06/2020 at https://www.cbs.nl/-/media/imported/documents/2014/44/2014-b57-pub.pdf.

NSB (2011). *The Irish Statistical System: The Way Forward and Joined Up Government Needs Joined Up Data National Statistics Board.* Government of Ireland. Last accessed on 18/06/2020 at https://www.nsb.ie/media/nsbie/pdfdocs/NSB_ISS_Position_Papers.pdf.

NSB (2015). *A World Class Statistical System for Ireland.* Government of Ireland. Last accessed on 18/06/2020 at http://www.nsb.ie/media/nsbie/pdfdocs/NSB_Strategy_2015-2020.pdf.

Ó Gráda, C. U. C. D. (2000). The Political Economy of the Old Age Pension : Ireland c. 1908- 1940. Last accessed on 18/06/2020 at http://www.ucd.ie/economics/research/papers/2000/WP00.22.pdf.

ONS UK (2013). Beyond 2011: Matching Anonymous Data. Technical Report July 2013. last accessed on 17/06/2020 at https://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-matching-anonymous-data–m9-.pdf.

ONS UK (2017). ONS Census Transformation Programme Annual Assessment of ONS ' s Progress Towards an Administrative Data Census. Technical Report June. Last accessed on 18/06/2020 at https://www.ons.gov.uk/census/censustransformationprogramme/ administrativedatacensusproject/administrativedatacensusannualassessments.

O'Sullivan, L. (2015). Linking, selecting cut-offs, and examining quality in the Integrated Data Infrastructure (IDI). *Statistical Journal of the IAOS*, 31(1):41–49.

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

Scholz, R. and Kreyenfeld, M. (2016). The Register-based Census in Germany: Historical Context and Relevance for Population Research. *Comparative Population Studies*, 41(2):175–204. Last accessed on 18/06/2020 at https://www.comparativepopulationstudies.de/index.php/CPoS/article/view/223/229.

Schwarz, N. (2001). The German Microcensus. *Schmollers Jahrbuch*, 121(2001):649–654. last accesed on 18/06/2020 at https://www.ratswd.de/download/schmollers/Schwarz.pdf.

Schwyn, M. and Kauthen, J.-p. (2009). The Swiss Census 2010: Moving Towards a Comprehensive System of Household and Person Statistics. *Insights on Data Integration Methodologies: ESSnet-ISAD workshop, Vienna, 29-30 May 2008*, (May 2008):110–123.

Statistics Act, e. I. S. B. (1993). Statistics Act. Last accessed on 18/06/2020 at http://www.irishstatutebook.ie/eli/1993/act/21/enacted/en/print.html.

Statistics Canada (2015). Census Technical Report : Coverage Census 2011. Technical report, Statistics Canada. Last accessed on 21/06/2020 at http://www12.statcan.gc.ca/census-recensement/2011/ref/guides/98-303-x/98-303-x2011001-eng.pdf.

Statistics Netherlands (2016). Usual Residence Population Definition : Feasibility Study The Netherlands. Technical report. Last accessed on 21/06/2020 at https://www.cbs.nl/-/media/_pdf/2017/08/statistics-netherlands-feasibility-study.pdf.

Statistics New Zealand (2012). Transforming the New Zealand Census of Population and Dwellings: Issues, Options, and Strategy, Wellington, New Zealand. Technical report. Last accessed on 21/06/2020 at https://www.stats.govt.nz/assets/Research/Transforming-the-New-Zealand-Census-of-Population-and-Dwellings-Issues-options-and-strategy/transforming-the-new-zealand-census-of-population-and-dwellings-issues-options-and-strategy.pdf.

Statistics New Zealand (2014a). Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey. Technical report, Statistics New Zealand. Last accessed on 21/06/2020 at http://archive.stats.govt.nz/˜/media/Statistics/browse-categories/population/census-counts/report-on-2013-post-enumeration-survey/report-on-2013-post-enumeration-survey.pdf.

Statistics New Zealand (2014b). Estimated Resident Population 2013 : Data Sources and Methods. Technical report, Statistics New Zealand. Last accessed on 21/06/2020 at http://archive.stats.govt.nz/˜/media/Statistics/browse-categories/population/estimates-projections/erp-2013-sources-methods/est-res-pop-2013-data-methods.pdf.

Statistics New Zealand (2016). Experimental Population Estimates from Linked Administrative Data : Methods and Results. Technical report. Last accessed on 21/06/2020 at http://archive.stats.govt.nz/˜/media/Statistics/surveys-and-methods/methods/research-papers/topss/exp-pop-estimates-linked-admin-data-methods-research/exp-popln-estimates-from-linked-admin-data.pdf.

Thygesen, L. (2010). The Importance of the Archive Statistical Idea for the Development of Social Statistics and Population and Housing Censuses in Denmark. Technical report, Statistics Denmark. Last accessed on 21/06/2020 at http://www.dst.dk/extranet/staticsites/Nordic2010/pdf/bf7d6701-5b9f-4888-adc2-a45ce8debf87.pdf.

Tiit, E.-M. (2014). 2011 Population and Housing Census Methodology. Technical report, Statistics Estonia, Tallinn.

Tønder, J.-K. (2008). The Register-based Statistical System. Preconditions and Processes. In *International Association for Official Statistics Conference*, Shanghai. Last accessed on 21/06/2020 at http://www.iaos-isi.org/papers/CS_1_2_Tonder.doc.

UNECE (2006). *Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing.* United Nations.

UNECE (2007). *Register-based Statistics in the Nordic Countries. Review of Best Practices with Focus on Population and Social Statistics.* United Nations.

UNECE (2008). *Measuring Population and Housing, Practices of UNECE Countries in the 2000 Round of Censuses.* United Nations.

UNECE (2014). *Measuring Population and Housing Practices of UNECE Countries in the 2010 Round of Censuses.*

UNECE (2015). *Recommendations for the 2020 Censuses of Population and Housing Conference of European Statisticians.* United Nations.

van der Heijden, P. G. M., Smith, P. A., Cruyff, M., and Bakker, B. (2018). An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates , with an Application to Mode of Transport of Serious Road Casualties. *Journal of Official Statistics*, 34(1):239–263.

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

Zhang, H. (2009). A Note About the Maximum Likelihood Estimator in the Hypergeometric Distribution. *Comunicaciones en Estadística*, 2(2):169–174.

Zhang, L.-C. (2011). A Unit-Error theory for Register-based Household Statistics. *Journal of Official Statistics*, 27(3):415–432.

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., 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. (2004). The Multiple-record Systems Estimator when Registrations refer to Different but Overlapping Populations. *Statistics in Medicine*, 23(14):2267–2281.