An Phríomh-Oifig Staidrimh
Central Statistics Office

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

INTERNATIONAL CONTEXT AND METHODOLOGY

Compiled by John Dunne, CSO, Ireland

Released as background material to
talk entitled "Demography Statistics - looking to the future"
as part of 7th CSO Administrative Data Seminar
"First Steps Towards a Virtual census"
hosted on 4th December 2018

Note -

This methodological note is an extract from project documentation associated with the PECADO research project. The note is laid out in two chapters. Chapter one, first, provides international context with respect to census modernisation efforts, before outlining a methodology to underpin a new system of population estimates at State level compiled from administrative data only. The second chapter then explores the robustness of the proposed system of population estimates. An extensive reference list is also provided.

Contents

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) on 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 high and 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 EUROSTAT (2003) published in 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. In a subsequent review in 2015 EUROSTAT (2015), Eurostat 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 (page 2) 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 page 2) 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.

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 report 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.

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 areas 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, organisational challenges and timeliness. Many countries looked to the experience of the 4 countries that did not have to contact each household and note 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 (NSI's). NSI's 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 is typically obsolete and needs 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 Denmarks 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 UNECE (2014), 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. 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 UNECE (2014) 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 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 reliable population estimates can be compiled directly from administrative registers, the associated estimates of population flows will also be reliable. The flows are simply identified and counted by comparing the population register at two points in time. Accurately identifying deaths and births will then also allow for reliable estimates of migration flows (emigration and immigration).

In the absence of suitable administrative registers, the compilation of reliable population estimates rely 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 a review of methods for estimating migration Jensen (2013), different approaches are considered 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.

The UNECE surveys UNECE (2014, 2008) 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 without having 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 follow.

- Census modernisation - mature

  Those countries that have conducted their Census enumerations directly from administrative sources or registers for a significant number of year. Examples of these countries are the Nordic countries and 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. Example countries include Israel, Spain, Austria and Germany.

- Census modernisation - aspiring

  Those countries that recognise the potential of administrative data sources and

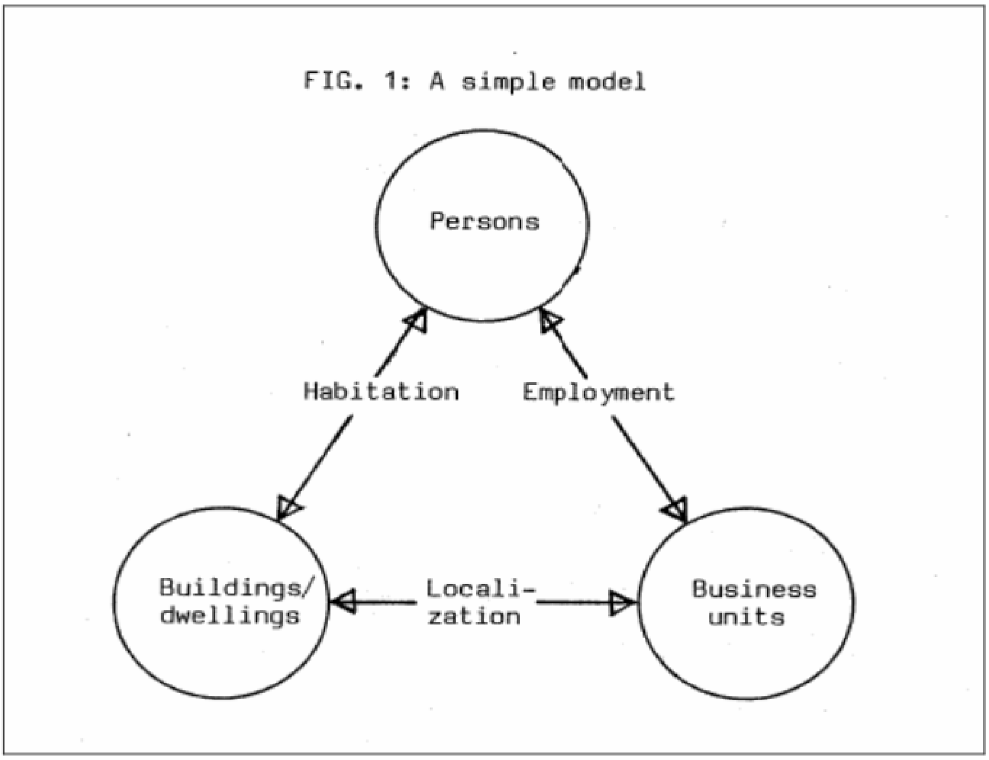

Figure 1.1: Representation of socio-demographic statistical system presented by Thygesen in early 80s Thygesen (2010)

are actively investigating systems and methods to conduct a Census for the first time with trying to contact every household to enumerate every person. Example countries include UK and New Zealand.

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, page 6. In fact, this 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 re-enforced 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 UNECE (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 Thygesen (2010) and Lange Lange (2014) provide an interesting history and understanding of this system developed in Denmark, including the many difficulties that were surmounted along the way. In particular, Thygesen 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 Lange (2014), the Census results are compiled by only two persons. This is predicated in already having the data collected and properly organised as required in a fully functioning register based statistical system.

Nordbotten 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. Krauss documents this proposal and the privacy concerns raised with a view to informing future policy decisions at the US Census Bureau Kraus (2010). The privacy concerns raised then are still relevant today.

Statistics Sweden recognise 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 previous 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 overcoverage.

Netherlands

As documented by Nordholt 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 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 is 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 use 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 state 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 do 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 the combining the Israel 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 was then estimated through the use of a sample of addresses in selected small areas and comparing the list of persons from the selected addresses with the list of persons in the IAF registered as living at that address. Undercoverage in the IAF for that small area is then generally estimated by identifying persons living in the selected area but having an address recorded as elsewhere in the IAF. Overcoverage in the IAF for the selected area 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 small areas are now used to estimate undercoverage and overcoverage rates in similar small areas in the rest of the country. 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 assumptions underpin the approach

- No erroneous enumeration in the survey
- Independence between the IAF and survey
- All persons have equal probability of being listed in the IAF in the correct area
- All persons have equal probability of being enumerated in the field
- Distribution of overcoverage in the selected small areas is proportional to the distribution of overcoverage in the population for the areas the selected small areas represent

This dual list method in 2008 depend on formal address systems which only cover 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 and undocumented residents or workers, 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 are also planning to include SoL type data for improving the local coverage estimation in the Census.

Spain

Spains 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 Estadistica, Spain, (INE) conducted a population and housing Census based on registers and a 10% sample surveyArgüeso 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 approximate 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 records as *sure*. 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:

$$CF_i = \frac{\hat{d}_i}{\hat{s}_i}\frac{\hat{S}_i}{\hat{D}_i}$$

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 pre-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

$$\hat{T}_i = S_i + CF_i D_i$$

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.

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 number or building identification number. 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 Scholz and Kreyenfeld (2016) were a significant consideration in the decision to conduct a register based Census in 2011. Germany did not conduct a Census in the 2000 round.

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 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 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 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 building 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 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 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: UK and New Zealand

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 is 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, date of birth to create a population spine. O Sullivan 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).

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 focussed 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), they highlight the need to use a new legal framework, 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, data 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.

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). Tnder 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 Act 2007 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 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

Every person being enumerated in the dwelling of residence is important to enabling cross classification of 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 as possible for the official population Nordholt et al. (2014). In the 2011 Spanish Census Argüeso and Vega (2014), where a 10% sample is 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.

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 translated 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 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 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 from 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), proposes 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 be described as linkage errors, domain misclassification errors and or coverage errors. Linkage errors will occur where there is an absence of official identifiers to enable high quality deterministic linkage, say in 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 Switzerland FSO (2015), Germany Bechtold (2016), Israel Kamen (2005) and Spain Argüeso and Vega (2014).

1.2.2 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 decentralised 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 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 in 2014 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.

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 on 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 has not had 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 be able to link across multiple data sources based on address. In the absence of this postcode system, the NDI faces the challenges 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 persons 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. 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).

1.2.3 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 takes the same role as a population frame created from the CPR.

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

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

One of the first milestones in moving to a register based 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 The Simple Idea

1.3.1 If You Don't Have a CPR, Build a Statistical Register

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 over coverage 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 there is no reason why linkage error cannot be 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 to be aware of is one of 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 to age with respect to 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 O Grada 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, group level or by prioritising data sources.

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. The SOL 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.* 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 years 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 includes 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 evidence of that person being resident in the State in 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 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 (M_t) in Wolter's 1986 paper Wolter (1986) is the starting point for our consideration of a model to adjust for undercoverage.

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

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

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.

		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 1.2: Multinomial distribution ϕ_i as per the Wolters *Multinomial Assumption* Wolter (1986)

		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 1.3: Cell counts from N mutually independent trials under Wolters *Autonomous Independence Assumption* Wolter (1986)

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 modelled by the multinomial distribution ϕ_i with parameters in table 1.2 (page 29).

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 1.3 (page 29) are considered observable. Cell count x_{22} and population size N are unknown and need to be estimated based on the model.

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** Nonrespondents can be identified such that exact matching between list A and list B can occur.

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 1.1 (page 30).

$$\theta_i = \frac{p_{i11}p_{i22}}{p_{i12}p_{i21}} = 1, \text{ for } i = 1, ..., N \tag{1.1}$$

9. Homogeneous Capture within List Assumption The capture probabilities satisfy $p_{i1+} = p_{1+}$ and $p_{i2+} = p_{2+}$ for $i = 1, ..., N$. (This assumption is numbered 11 in Wolters paper and not labelled.)

Under these assumptions and using Maximum Likelihood methods, Wolter shows that an estimator for N is given by equation 1.2 (page 30) with a variance estimator given by equation 1.3 (page 30)

$$\hat{N} = \frac{x_{1+}x_{+1}}{x_{11}} \tag{1.2}$$

$$\hat{V}\left[\hat{N}\right] = \frac{x_{1+}x_{+1}x_{12}x_{21}}{x_{11}^3} \tag{1.3}$$

In considering Wolters 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 Wolters assumptions (*5. Spurious Events Assumption*, *6. the Nonresponse Assumption.* and *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 Wolters *Multinomial* and *Autonomous Independence Assumptions.*

1.3.2.3 Dual System Estimation Methodology - Adjusted

We take as our starting point Zhang and Dunne 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 $i \in (B \cap U)$, 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 $i \in A \cap U$, 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 $\boldsymbol{\delta}_A = (\delta_{1A}, ..., \delta_{xA})$ as fixed constants, where $\sum_{i \in A} \delta_{iA} = x$. 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:

$$E(m|\boldsymbol{\delta}_A) = \pi$$

which is the expectation of the number of records in list AB 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 method of moment estimator, given by

$$\hat{N} = nx/m \tag{1.4}$$

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 assumption are more relaxed than those described in Wolters 1986 paper Wolter (1986). With respect to Wolters 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. 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

$$\begin{aligned} V(\hat{N}) &\approx \frac{x^2}{E(m)^2}\left(V(n) - \frac{2E(n)}{E(m)}Cov(n,m) + \frac{E(n)^2}{E(m)^2}V(m)\right) \\ &= N(\frac{1}{\pi} - 1)\left(\frac{N}{x} - 1\right) \end{aligned}$$

Replacing N by xn/m and π by m/x, we have

$$\hat{v} = \widehat{V}(\hat{N}) = \frac{n(n-m)x(x-m)}{m^3} \tag{1.5}$$

Notice that this is the same variance estimate as that of the standard DSE described in the text book *Discrete Multivariate Analysis* 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 Wolter (1986) need to apply is weak.

Chao et al 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.

1.3.2.4 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. Those that do not hold a driving licence typically have the same right to apply for, and hold, a driving licence as those that do hold a driving licence (typically a provisional/learner driver licence for first time applicants).

The list of those persons that renewed their driving licence or applied for a new one in the relevant 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, in recent years 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 Driving 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.

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.

Under the closed population assumption,

$$U_A = U_B = U_{II} \tag{1.6}$$

that is the hypothetical SPD population equates to 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.

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 drivers 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.

The next chapter explores this system of population estimates to see how robust it is. In particular, we introduce the concept of Trimmed Dual System Estimation (TDSE) as a 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 look to the population estimates obtained from the Census of Population 2011 and compare with the proposed new system of population estimates.

Chapter 2

Is the PECADO system of population estimates robust?

2.1 Introduction

In this chapter we explore the system of population estimates proposed in section 1.3 (page 24).

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.

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, it 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 as in the PPSN.

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 2.2, we present the population estimates and consider the availability of the underlying data sources. Section 2.3 considers the homogeneous capture assumption. Then section 2.4 considers the presence of erroneous records and presents a new toolkit (Trimmed Dual System Estimation - TDSE) that allows statisticians to hunt for overcoverage. The penultimate section, section 2.5 will then consider how dependent the system of population estimates on each underlying data source.

Section 2.6 concludes the chapter with a summary of our findings and insights with respect to the proposed system of population estimates.

2.2 A system of population estimates

2.2.1 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 in more detail the dependence of this system on individual data sources later in section 2.5 (page 56).

Table 2.1 (page 39) provides a summary of the availability of the different data sources by calendar year.

Figure 2.1 (page 39) illustrates the coverage of each source for the chosen reference year with respect to the PAR.

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
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 2.1: Availability of Data Sources by year. QNHS is further described in section 2.3.2 (page 43)

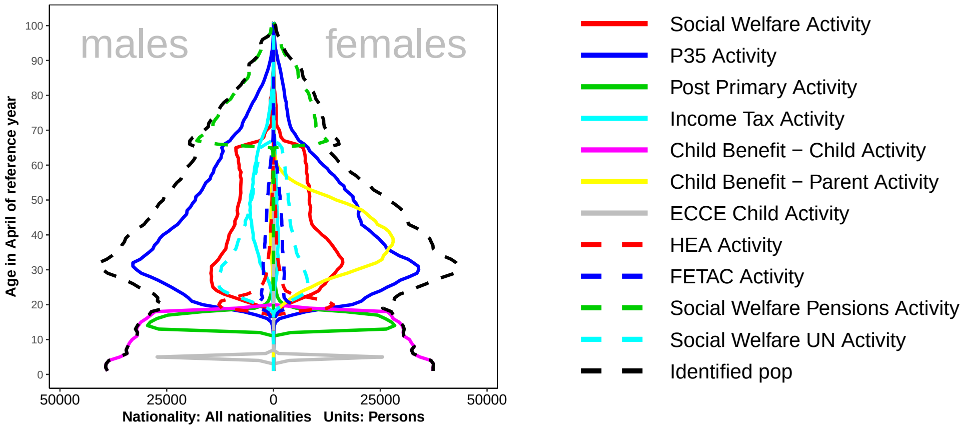

Figure 2.1: Administrative Data Source Coverage with respect to PAR, 2011

In looking at the school age and pre school age part of the population tree in figure 2.1 (page 39), 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

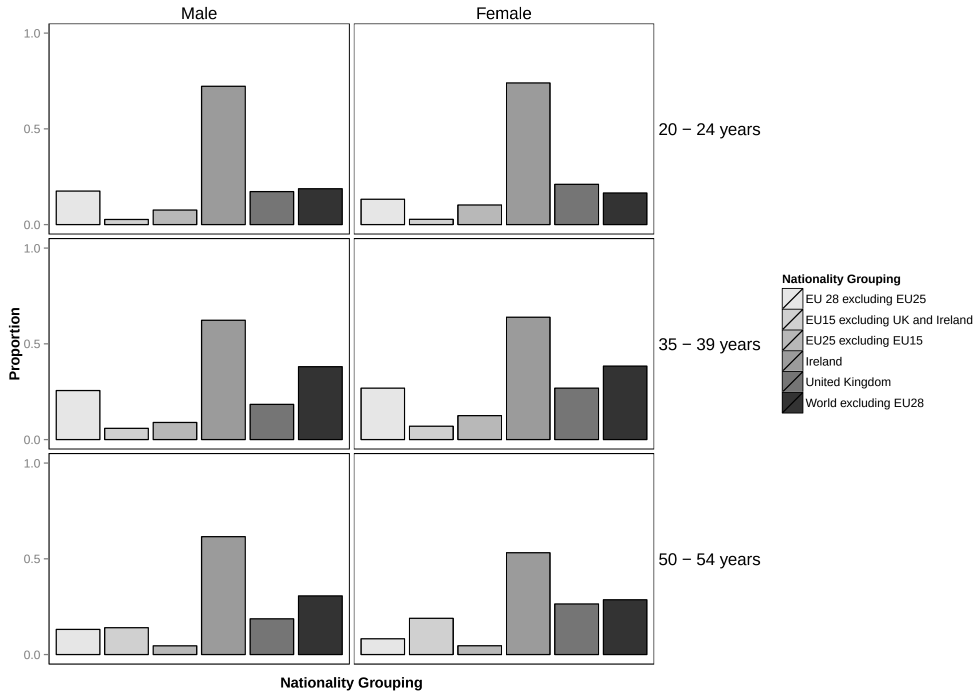

Figure 2.2: Proportion of identified Irish driver licence holders on PAR by nationality group, selected age group and sex, 2011 - REDO in COLOUR

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 is available in later years is a data source related to public health care (PCRS), this 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 2.2 (page 40) 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. 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.

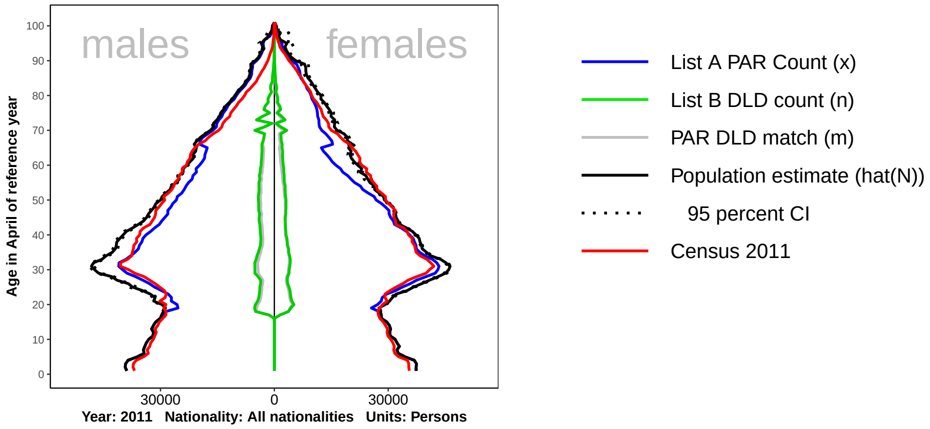

Figure 2.3: Preliminary Population Estimates by sex and single year of age, All nationalities, 2011

While not every person in the resident population will hold a drivers licence or apply for a drivers licence, every person resident in the population is entitled to apply for a driver licence and as such has a probability of being included in list B (DLD). 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. We will examine the strength of this assumption later in section 2.3 (page 43).

2.2.2 Population estimates

We now adjust the SPD counts using the DSE methodology outlined in section 1.3.2.3 (page 31) to obtain a first set of population estimates and present them in figure 2.3 (page 41). The figure presents the population estimates along with the SPD and DLD counts using a population tree 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 also plotted using dotted lines but the relative size of the confidence intervals is such it is nearly 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.
- 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 2.1 (page 39) 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 substituted in as a list B in the proposed system, this second source will also help to validate the homogeneous capture assumption for DLD as a list B.

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

2.3.1 Evaluation strategy

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 similar. 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.

2.3.2 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

- Survey design sample of 25,000 households per quarter
- Two stage design, where each household has equal probability of being selected
- Survey achieved sample 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.

- 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.

More information on the QNHS is available at http://www.cso.ie/en/qnhs/qnhsmethodology/ (accessed on 15th August 2017).

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. This results in a list B with approximately 60,000 persons (just over 1% of the population) that can be assumed to be usually resident in the State in a 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. 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.

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.

2.3.3 DLD V QNHS - a list B comparison

Figure 2.4 (page 45) 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 to the survey and the fact that if one person in a house is captured all individuals in that house will be captured for the survey violates the assumption of independent capture for individuals in estimating the variance as in equation (1.5). Therefore, the confidence intervals in practice will be wider than those shown for the QNHS based estimates.

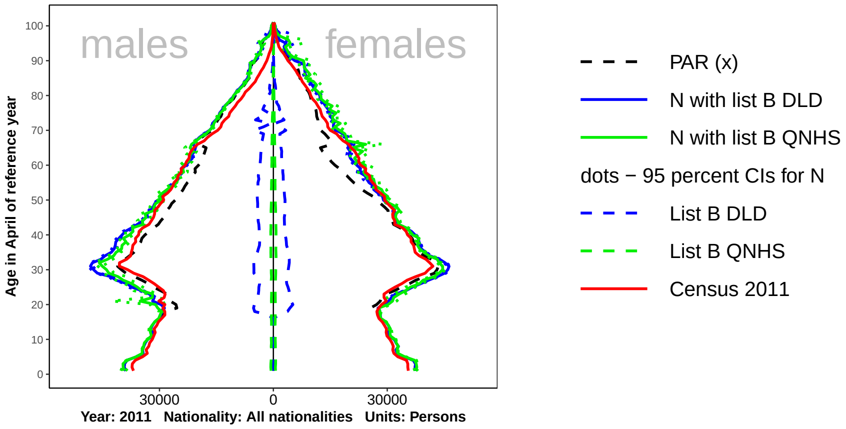

Figure 2.4: Comparison of Population Estimates under two different List B's, All nationalities, 2011

In comparing the two sets of figures there is very little difference across age group and gender. The only potentially significant differences occur where the population estimates peak for both males and females at age 30. 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.

A number of explanations could be considered to explain 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 that to renew a drivers licence a person must reside in the State. 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 a plausible explanation as there is a cost to letting a drivers licence lapse - a person may have to resit their driving test.

However, this argument could also be reversed in to 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. If these groups can be identified with subsets of persons within blocks, then 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 at figure 2.4 (page 45) 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 for this age group.

2.4 DSE with erroneous records in list A Trimmed Dual System Estimation (TDSE)

2.4.1 Ideal DSE, given erroneous enumeration

We take as our starting point the DSE methods as described in section 1.3.2.3 (page 31). 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 subjected 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 subjected 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 subjected 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 assumption of *matching*. Suppose that error-free matching between A and B gives rise to the matched list AB with m records.

Let $\delta_{iB} = 1$ if $i \in B \cap U$, 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).

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 $\boldsymbol{\delta}_A = (\delta_{1A}, ..., \delta_{xA})$ as fixed constants, where $\sum_{i \in A} \delta_{iA} = x - r$.

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

$$\tilde{N} = n(x - r)/m \tag{2.1}$$

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 $r > 0$.

Again the estimator (2.1) is based on the method of moment instead of maximum likelihood.

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 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.

2.4.2 Trimmed DSE

The estimator (2.1) is hypothetical because one does not actually know r, i.e., the number of erroneous records in list A. But one *can* (a) score some records in list A, which are most likely to be erroneous, (b) match them to list B and, then, (c) calculate the DSE as if list A would have been free of erroneous enumeration once the scored records had been removed.

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

$$\hat{N}_k = n\frac{x - k}{m - k_1} \tag{2.2}$$

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

The trimmed DSE can be compiled under the *same* assumptions as those for the ideal DSE, as given by equation (2.1), *regardless* of how systematic the scoring is in removing the records in list A, for the same reason that potential systematic under-coverage of list A does not matter to start with. For instance, had one scored 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.

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 result is useful.

Result 1: If $k_1/m < k/x$, then $\hat{N}_k < \hat{N}_0$. If $k_1/m = k/x$, then $\hat{N}_k = \hat{N}_0$.

Now that mk/x is the expectation of k_1 under *random* scoring of k out of the x records in list A, Result 1 implies that one can expect the trimmed estimate, equation (2.2), to be lower than the naïve estimate nx/m, as long as a relatively smaller number of scored records are confirmed to be non-erroneous, because they can be 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.

Meanwhile, $\tilde{N}$ would be the optimally trimmed estimate with $(k, k_1) = (r, 0)$. It seems desirable to avoid 'over-trimming' that makes the trimmed DSE, equation (2.2), lower than the ideal DSE, equation (2.1).

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

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 $k_1/m \geq 0$. □

In other words, Result 2 assures that over-trimming will not be the case, as long as one does not score more records than the number of erroneous records that exist in list A. For instance, if it is suspected that about 10% of the records in list A may be erroneous, then over-trimming can be avoided as long as one does not score more than 10% of the records in list A.

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

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

To summarise, as long as one is able to score the erroneous records in list A more effectively than random scoring *and* one does not score more records than the total number of erroneous records in list A, the trimmed DSE, equation (2.2), can be expected to reduce the bias of the naïve DSE and move it closer to the ideal DSE, equation (2.1). Provided the scoring has succeeded in removing all the erroneous records, the expectation of the trimmed DSE would 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

$$\begin{aligned} V(\tilde{N}) &\approx \frac{\tilde{x}^2}{E(m)^2}\left(V(n) - \frac{2E(n)}{E(m)}Cov(n,m) + \frac{E(n)^2}{E(m)^2}V(m)\right) \\ &= N(\frac{1}{\pi} - 1)\left(\frac{N}{\tilde{x}} - 1\right) \end{aligned}$$

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

$$\tilde{v} = \widehat{V}(\tilde{N}) = n(n-m)\tilde{x}(\tilde{x}-m)/m^3$$

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 scored list A with x_k records, which is not possible. 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

$$v_k = \widehat{V}(\hat{N}_k) = n(n-m_k)x_k(x_k-m_k)/m_k^3$$

2.4.3 Stopping rules

Not withstanding 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 the scoring 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 scoring becomes more or less random at picking out the 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 repeat the scoring to successively include more records, and to keep track of the actual $\hat{N}_k$, as k increases, to see if it flattens out at some stage.

Secondly, it is intuitive that k_1, i.e. the number of scored records that are confirmed to be non-erroneous, should be as low as possible. Denote by p the probability that a scored record is actually erroneous. Let $k_r = r/p$ be the expected number of records, in order to score 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 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 scoring is at picking out the erroneous records, the bigger the change. On the other hand, in the case of random scoring or worse, we would have $p \leq r/x$ and $k_r \geq x$. Since it is not possible to score more than x records in list A, one cannot expect to detect any change in the ratio k_1/k with any such scoring method.

It should be pointed out that, in reality, it is unlikely that the probability p of scoring erroneous records will be a constant of k, i.e. the number of records scored. However, the above consideration suggests that, in practice, one could repeat the scoring to successively include more records, and to keep track of the actual k_1/k, 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 scored, k_1/k may be roughly convex, in which case the stopping point could be where the bend is most acute.

Thirdly, because the way in which 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.

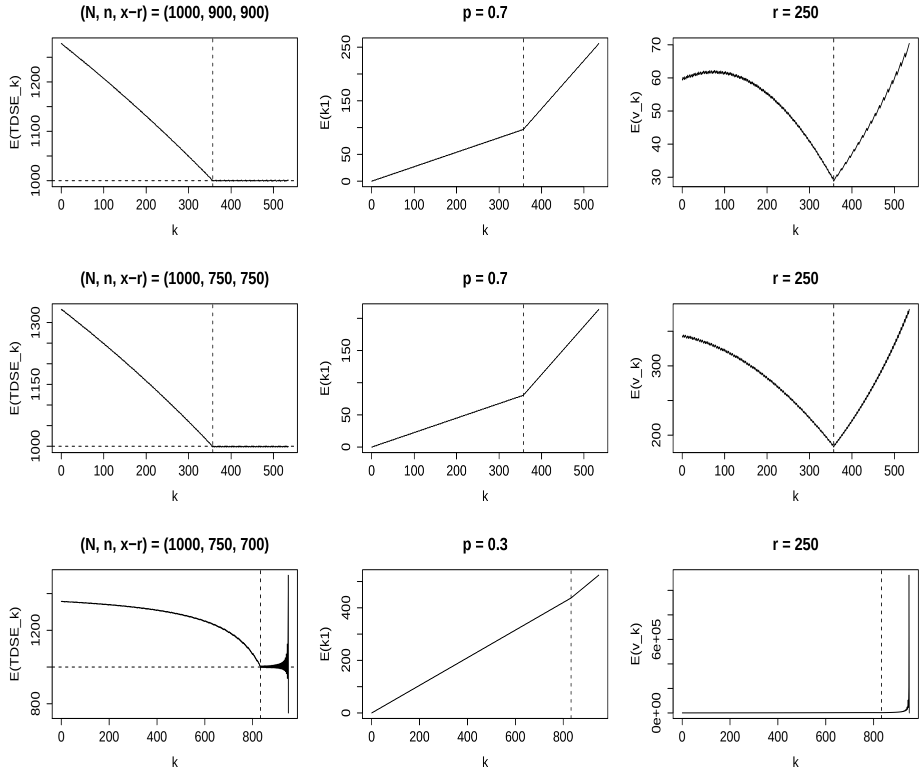

Figure 2.5: Illustration of three stopping rule indicators using simulated data: left column: $E(\hat{N}_k)$; middle column: $E(k_1)$; right column: $E(v_k)$. Setting (N, n, x, r, p): same for each row.

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

Figure 2.5 (page 51) provides the illustration using simulated data.

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 scoring erroneous records, an unfavourable scenario with both low π and p, and a scenario between these with low π but reasonably high p.

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 scoring 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 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 high scoring 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 random scoring (at the rate 250/950) in this case. Removing all the erroneous records at such a rate requires on expectation scoring 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.

In summary, the performance of trimmed DSE is above all determined by how effectively the scoring removes the erroneous records. The trimmed DSE can yield good bias-variance trade-off compared to the naïve DSE, even when a fair amount of records need to be removed from the estimator. Of course, in practice, it may be impossible to remove all the erroneous records by scoring, or one may lack very effective means of scoring. 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.

2.4.4 TDSE Applied

We now apply the TDSE to our system of population estimates to illustrate the method.

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 2.6 (page 53) presents 3 different situations (one per row) with respect to the stopping rules described in Section 2.4.3. 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.

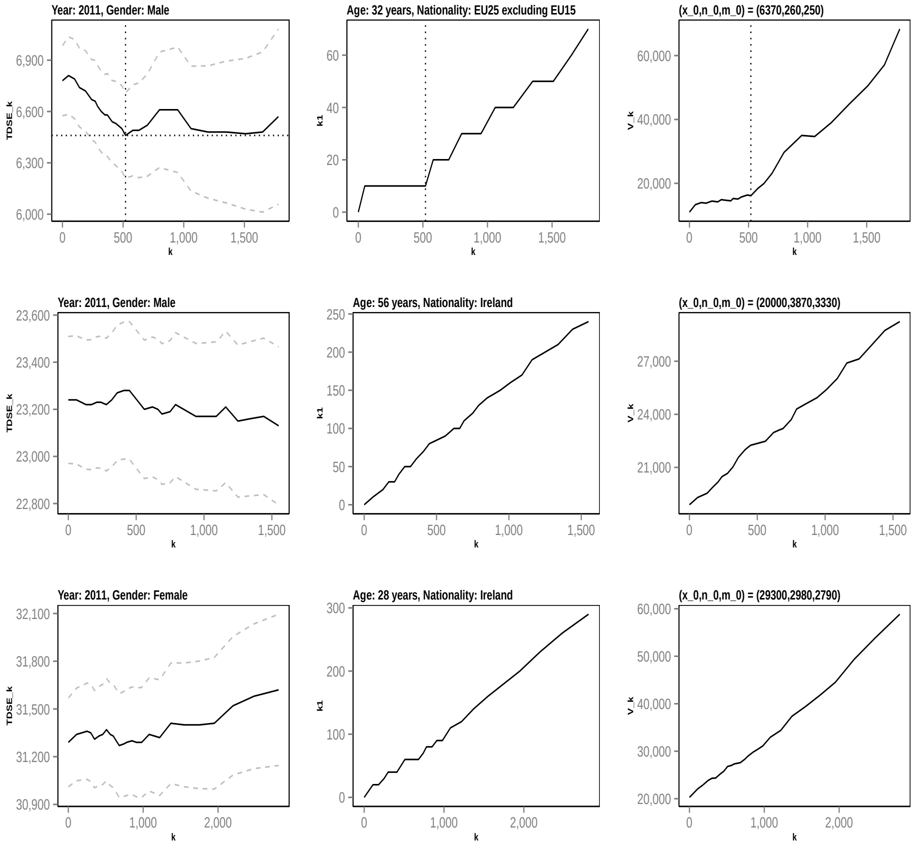

Figure 2.6: Illustration of TDSE in year 2011. Left column: TDSE $\hat{N}_k$ with 95% CI; middle column: k_1; right column: $V(\hat{N}_k)$. Each row presents a different population post-stratum. First row: Males aged 32 years with a nationality EUnew; Second row: Males aged 56 years with an Irish nationality; Third row: Females aged 28 years with an Irish nationality. All figures are rounded to nearest 10.

More explicitly, the first stopping rule looks to see if $\hat{N}_k$ flattens out at some point, indicating that the scoring 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 scoring method is removing less erroneous records than would be the case if it were removing the records at random.

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 is most acute. The first case is the only one with a slightly convex curve with the bend being 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 once 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 scored 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 scoring method. Notice also that the difference between the TDSE and untrimmed DSE is not significant with regard to the 95% CIs.

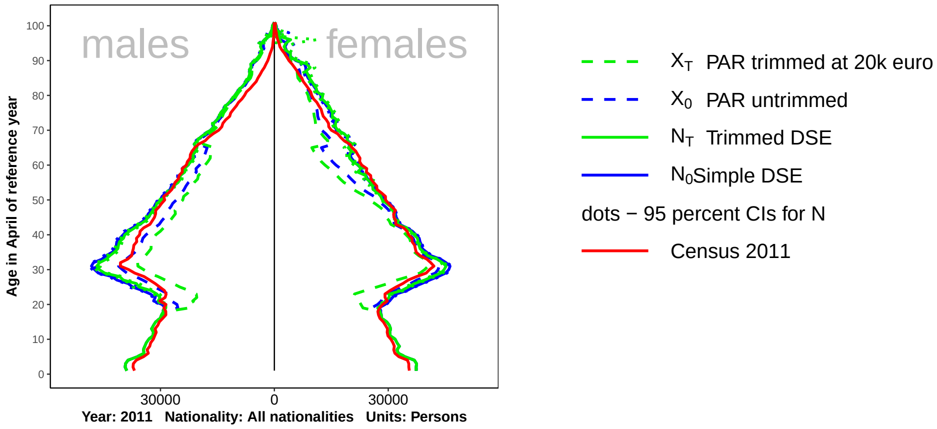

Figure 2.7: Population estimates with SPD trimmed at 20k euro by age and sex, 2011

For an appreciation of the overall effects of trimming, we refer to the population tree in Figure 2.7 (page 55) 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 scoring 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. In this instance the DSE is simply estimated as the X_0 or all those identified with activity. Where N_0 and 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 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 or slightly higher than the corresponding DSE. This is the age when many young people enter the work force and the number of scored 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.

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 in Ireland for 12 months or intend to reside in Ireland for 12 months. Figure 2.8 (page 56) shows the impact of this trimming to be almost negligible when considering all nationalities.

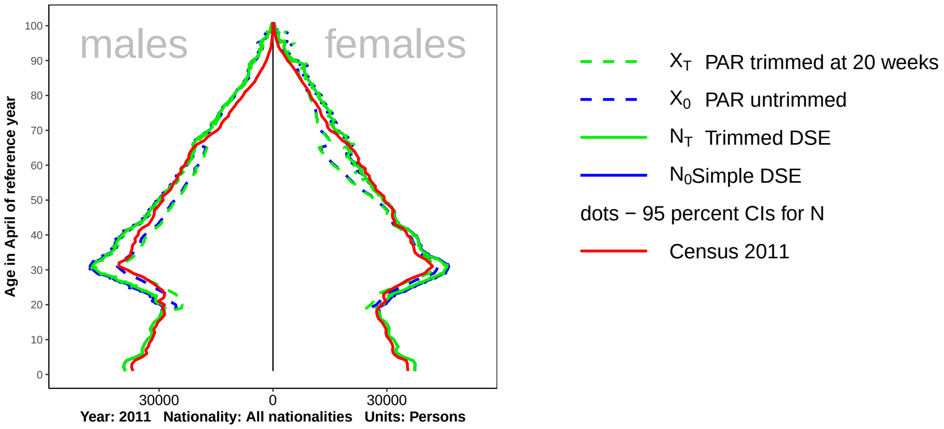

Figure 2.8: Population estimates with SPD trimmed at 20 weeks work by age and sex, 2011

2.5 Evaluation of individual data sources

2.5.1 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 population estimates 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 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.
  - 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 trees, similar to those earlier in the chapter to evaluate the impact of a given data source on the population estimates.

Using table 2.1 (page 39) and figures 2.1 (page 39) and 2.3 (page 41) 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 2.3 (page 41) 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.

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

Figure 2.9 (page 58) 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 if confidence intervals are considered to determine significance. This hints at the possibility of some small pockets of erroneous records in this data source.

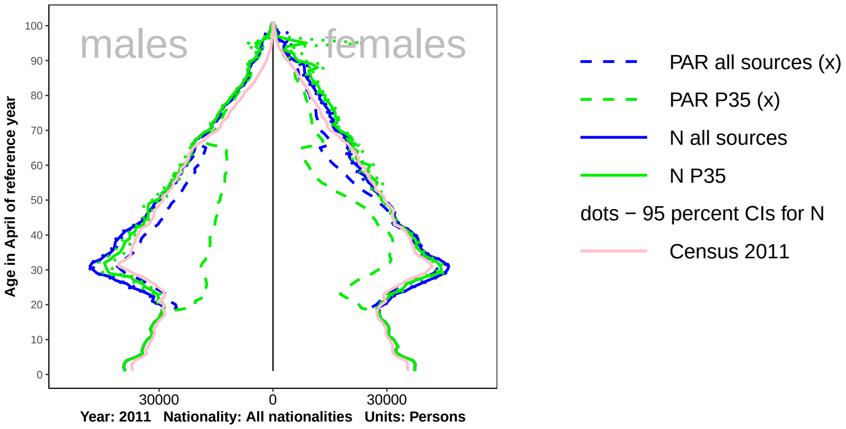

Figure 2.9: Population estimates with SPD trimmed of P35 - all employee records, List B = DLD, 2011

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.

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

Figure 2.10 (page 59) presents an analysis of what happens to the population estimates when P35 records are not included in the compilation of the SPD.

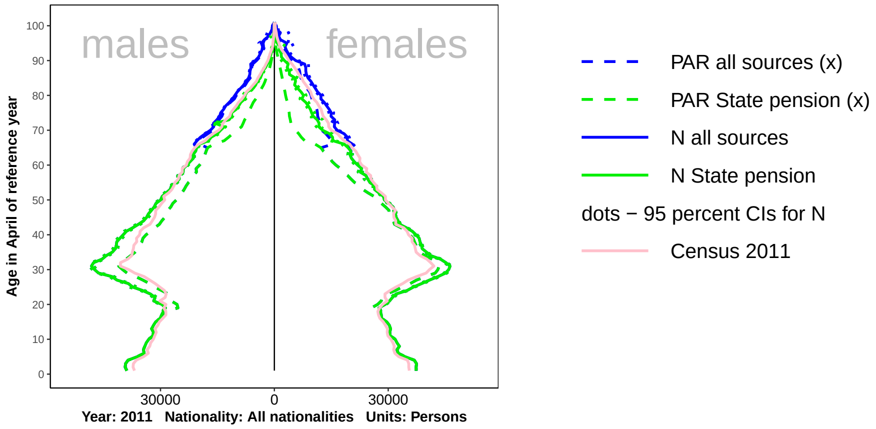

Figure 2.10: Population estimates with PAR trimmed of State pension records, List B = DLD, 2011

Trimming the state pension records from the SPD results in a sizeable 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. Given that the untrimmed SPD counts are highly unbalanced between the older males and females, 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. 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 2.3 (page 43), where the homogeneous capture assumption was evaluated. The presence of erroneous records in the

State pension source (with erroneous records) 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.

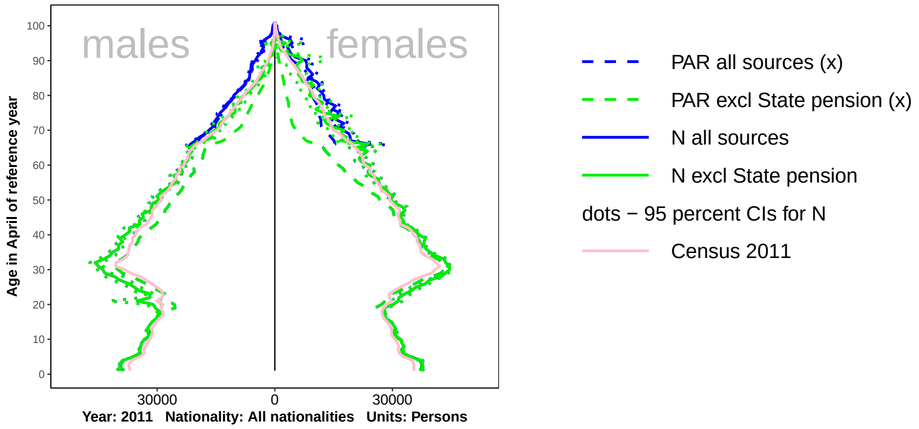

Figure 2.11: Population estimates with SPD trimmed of State pension records, List B = QNHS, 2011

Figure 2.11 (page 60) presents the same analysis as figure 2.10 (page 59) 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 accurate 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 Gerritse et al. (2016), whereby they demonstrate that when list counts are low, there is high sensitivity to the *independence assumption* (as described by Wolter Wolter (1986)), we can draw a parallel conclusion that when list counts are low there is high sensitivity to violations in the *homogeneous capture* assumption for list B. Therefore, in order to increase precision and accuracy in this age group, it is necessary 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 2.12 (page 61) 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.

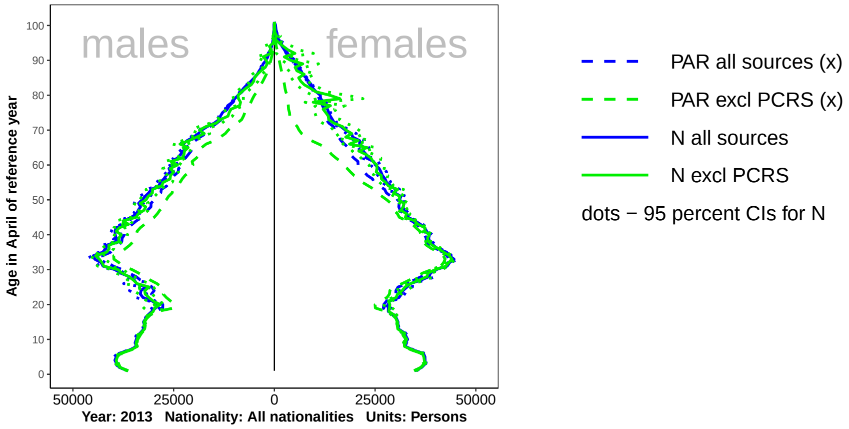

Figure 2.12: Population estimates, List B = QNHS, 2013. State pensions data source also excluded.

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

From figure 2.1 (page 39) it is obvious that it is impossible to estimate the population under 12 years of age without 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 2.5.3 (page 59) points to significant suspicion of erroneous records in this source. Figure 2.13 (page 62) presents this analysis again using the population tree 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 (PPPDB) 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 CB data source.

While the compilation of population estimates using the trimmed PAR for the under 12 age group is possible, the estimates have significantly wide confidence intervals such that they are not practical in use. 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.

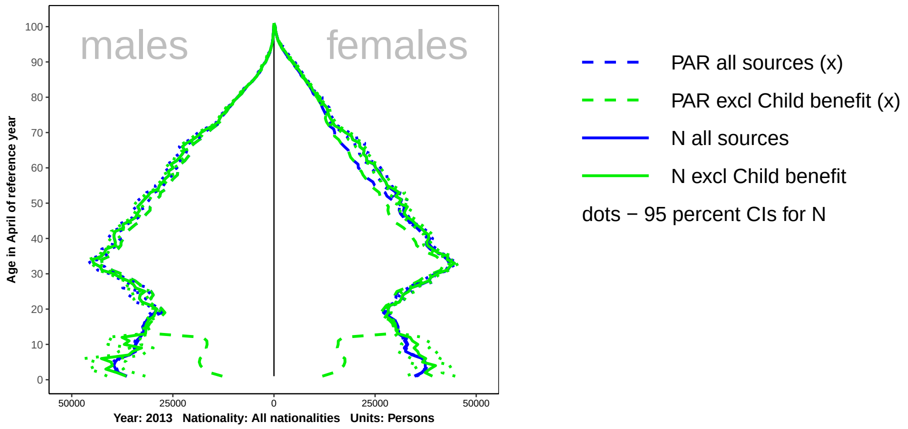

Figure 2.13: Population estimates with PAR trimmed of Childrens Benefit records, List B = QNHS, 2013. State pensions data source also excluded.

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.

2.6 Final reckoning

2.6.1 List B considerations

In section 2.3 (page 43) 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 2.3 (page 43) 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 satisfies the homogeneous capture assumption also. The set of estimates compiled using QNHS will not be as precise as those compiled using DLD due to the size of the datasets.

In a subsequent analysis presented in section 2.5.3 (page 59), we suspect State pension data source to contain considerable numbers of erroneous records. We also compile population estimates excluding State pension using two different data sources and present these in figure 2.10 (page 59) which uses DLD as list B and in figure 2.11 (page 60) 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. The presence of erroneous records in the State pension data source looks to have hidden this finding in the analysis in section 2.3 (page 43).

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 coverage is low as using DLD will not fully adjust for the undercount. When the PAR coverage is high, the adjustment required is small and as such using DLD instead of QNHS may not have such 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 found (SPD is considered as having complete coverage) then there is no requirement to adjust for undercoverage.

2.6.2 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. When this source was investigated further it was found that, in the absence of the actual payment data, a proxy indicator was used and it is believed the erroneous records are associated with the development of the proxy.

The precision of estimates the over 65 years age group suffer in the years prior to reference year 2013. In 2013 the PCRS data source first becomes 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 and the assumption is made that there has been no immigration in this age category for these 2 years.

PCRS is identified as a key data source with respect to 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.

Other adjustments to the PAR before finalising population estimates include preprocessing to ensure for each data source, where there is evidence that the person does not reside in Ireland, that those records are removed prior to compiling the SPD.

2.6.3 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.

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 into the final estimates compiled from the trimmed SPD and chosen list B.

The strategy used is described in figure 2.14 (page 65) 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.

2.6.4 Results

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

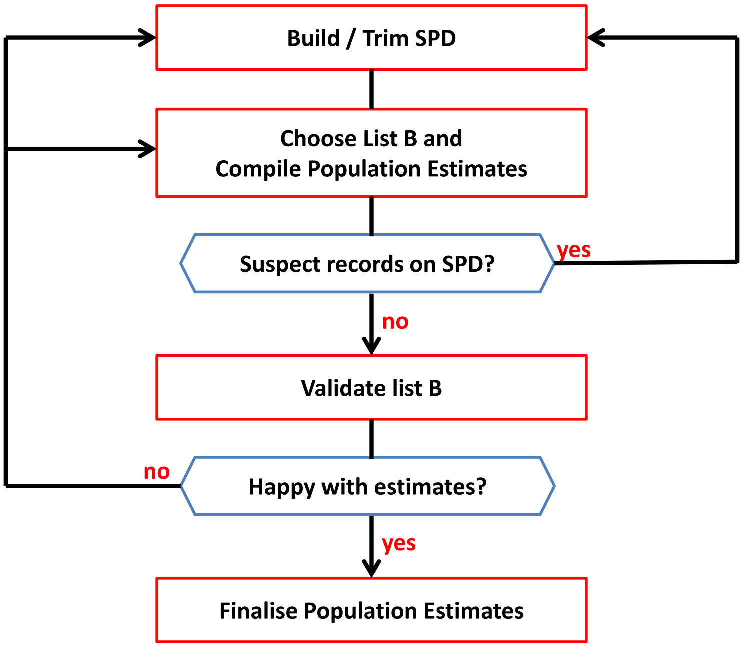

Figure 2.14: High level process map for compilation of population estimates.

- 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 for 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 positive dependence in using DLD for these categories will only have a minor impact on the bias. A positive dependence will lead to a negative bias in the estimate or an under estimate of the population.

- 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 concept of usual residence (12 months residence, intention or actual) that is commonly used.

The population estimates, along with precision estimates, for years 2011 to 2016 are presented in table 2.2 (page 67). 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 2.1, page 39, 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 2.3 and this comparison, broken down by age is presented using population trees in figures 2.15 (page 68) and 2.16 (page 68).

The gap between the new population estimate and the census usual resident account 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 trees 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.

Three possible explanations are considered to explain 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 count emigration prior to census night in the calendar year and immigration subsequent to census night in the calendar year must be added for usual residents.

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

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) (thousands)						
	422,680	507,030	468,870	378,100	466,610	539,200
Match between list A and list B (DLD) (thousands)						
	376,950	452,730	425,130	341,230	422,070	462,330
Population estimate (thousands)						
	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
Ref : Estimates published as Research Outputs December 2018 on http://www.cso.ie						

Table 2.2: Population estimates compiled from administrative data sources, 2011 to 2016

	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
Published on http://www.cso.ie December 2018 as Research Outputs				

Table 2.3: Comparison of Population Estimates with census usual resident counts by gender, 2011 and 2016

no erroneous records on DLD. The DLD has also been validated in section 2.3 (page 2) 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 has a significant quantity of erroneous records.

A third explanation is that a violation of the homogeneous capture assumption could lead to bias in the population estimates. While earlier analysis (see section 2.3, page 43) generally validated this assumption there were indications of a small positive dependence between the PAR and DLD for females in the older age category. This small positive dependence between DLD and PAR can be evidenced in figure 2.10 (page 59) 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 2.11 (page 60), the population estimate moves back above the Census estimate albeit with a larger confidence interval.

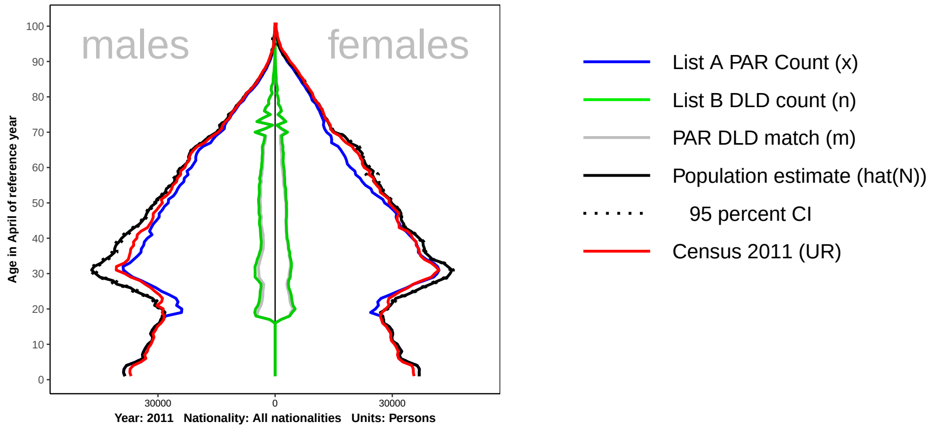

Figure 2.15: Comparison of final population estimates and Census usual resident counts by age and sex,2011

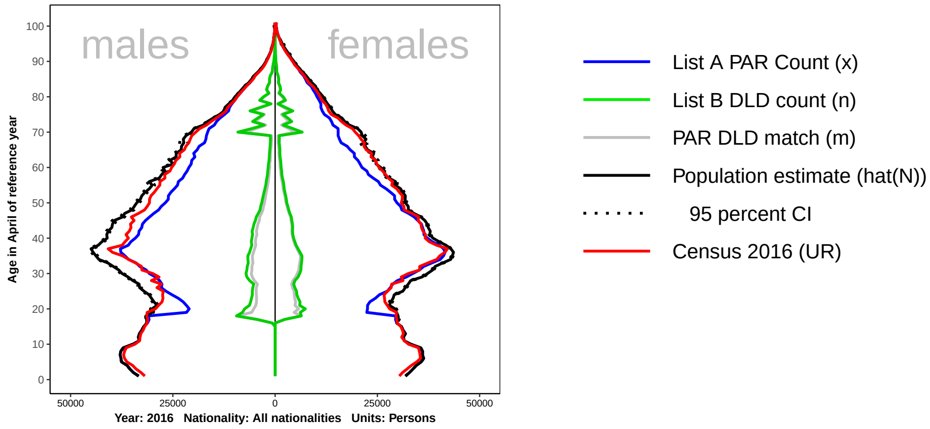

Figure 2.16: Comparison of final population estimates and Census usual resident counts by age and sex,2016

2.7 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 UNECE (2014). Lanzieri 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. In the absence of personal identification numbers, the system would then have to deal with the highly likely situation of erroneous records 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 is addressed. Zhang and Dunne Zhang and Dunne (2018) also consider linkage error in the DSE model.

To the authors knowledge, this is the first example of a system of national population estimates compiled solely form 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 to be applied 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 in negotiating access to activity or SoL data to compile population estimates.

References

Abbott, O. (2009). 2011 UK Census Coverage Assessment and Adjustment Methodology. *Population Trends*, 137(1):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.

Bechtold, S. (2016). The 2011 Census Model in Germany. 9:1–10.

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*, number October. UNECE.

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*, number October.

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

Chao, A., Pan, H. Y., and Chiang, S. C. (2008). The Petersen - Lincoln estimator and its extension to estimate the size of a shared population. *Biometrical Journal*, 50(6):957–970.

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.

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.

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.

DES (2013a). Early Leavers What Next? Report on Early Leavers from Post-Primary schools . Technical report, Department of Education and Skills.

DES (2013b). School Completers What Next? Report on School Completers from Post-Primary Schools . Technical report, department of Education and Skills.

DPER (2011). 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.

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.

EUROSTAT (2003). *Demographic statistics: Definitions and methods of collection in 31 European Countries.* European Communities.

EUROSTAT (2015). *Demographic Statistics: A review of definitions and methods of collection in 44 European countries.* Eurostat.

FSO (2015). New census system Quality survey. Technical Report January, FSO.

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 CensusesGroup of Experts on Population and Housing Censuses*, number Sepetmber, pages 1–15. UNECE.

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.

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.

INE Spain (2014). Population figures methodology. Technical Report July.

INE Spain (2018). Migration Statistics Methodology. Technical Report February.

Jensen, E. (2013). A Review of Methods for Estimating Emigration.

Kamen, C. S. (2005). The 2008 Israel integrated census of population and housing. *Statistical Journal of the United Nations ECE*, 22:39–57.

Kraus, R. S. U. C. B. (2010). Statistical Deja Vu: The National Data Center Proposal of 1965 and Its Descendants. Technical report.

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.

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

Macfeely, S. and Dunne, J. (2014). Joining up public service information: The rationale for a national data infrastructure. *Administration*, 61(4):93–107.

Mcnally, J. and Bycroft, C. (2015). Quality standards for population statistics : Accuracy requirements for future census models Census Transformation. Technical report, Statistics New Zealand.

Nordbotten, S. (2010). The statistical archive system 1960-2010: A summary. *Nordisk Statistikermøde i København*, 11.

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.

Nordholt, E. S., Van Zeijl, J., and Hoeksma, L. (2014). *Dutch census 2011: Analysis and methodology*. Statistics Netherlands.

NSB (2011). *The Irish Statistical System: The Way Forward and Joined Up Government Needs Joined Up Data National Statistics Board*. Government of Ireland.

NSB (2015). *A World Class Statistical System for Ireland*. Government of Ireland.

Ó Gráda, C. U. C. D. (2000). The political economy of the old age pension : Ireland c. 1908- 1940.

ONS UK (2013). Beyond 2011: Matching Anonymous Data. Technical Report July 2013.

ONS UK (2017). ONS Census Transformation Programme Annual assessment of ONS ’ s progress towards an Administrative Data Census. Technical Report June.

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 edit 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.

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.

Statistics Netherlands (2016). Usual Residence Population Definition : Feasibility Study The Netherlands. Technical report.

Statistics New Zealand (2012). *Transforming the New Zealand Census of Population and Dwellings: Issues, options, and strategy, Wellington, New Zealand.*

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

Statistics New Zealand (2014b). Estimated resident population 2013 : Data sources and methods. Technical report, Statistics New Zealand.

Statistics New Zealand (2016). Experimental population estimates from linked administrative data : methods and results.

Thygesen, L. (2010). The importance of the archive statistical idea for the development of social statistics and population and housing censuses in Denmark (Betydningen af den arkivstatistiske idé for udvikling af social statistik og folketaellinger). Technical report, Statistics Denmark.

Tønder, J.-K. (2008). The Register-based Statistical System. Preconditions and Processes. In *International Association for Official Statistics Conference*, Shanghai.

UNECE (2006). Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing. Technical report.

UNECE (2007). *Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics.*

UNECE (2008). *Measuring Population and Housing, Practices of UNECE countries in the 2000 round of censuses.* United Nations Publication.

UNECE (2014). Measuring population and housing. practices of UNECE countries in the 2010 round of censuses. Technical report, UNECE.

UNECE (2015). *Recommendations for the 2020 Censuses of Population and Housing Conference of European Statisticians.* United Nations.

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

Zhang, L. 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 Trimmed du, pages 237–258. CRC press.