To count or to estimate: a note on compiling population estimates from administrative data

John Dunne
Central Statistics Office (CSO), Cork, Ireland.
E-mail: John.Dunne@cso.ie
Francesca Kay
Central Statistics Office (CSO), Cork, Ireland.
Timothy Linehan
Central Statistics Office (CSO), Cork, Ireland.

Summary. Like many countries, Ireland has been researching new systems of population estimates compiled using administrative data. Ireland does not have a Central Population Register from which the estimates can be compiled.
The primary step in compiling population estimates from administrative data is to first build a Statistical Population Dataset (SPD). Ideally an SPD will have one record for each person in the population containing the relevant attributes. The ideal SPD then allows compilation of statistics by simply counting over records.
In practice, the compilation of SPDs is prone to error. These errors can be classified into 4 types of error; over coverage, under coverage, domain misclassification and linkage error. Ireland, to date, has investigated 2 different approaches to the compilation of population estimates from administrative data. The first, labeled in this paper as the *simple count method*, is based on building an SPD which minimises the overall number of individual record errors such that simple counts from the SPD will provide population estimates. The second, labeled in this paper as the *estimation method*, is based on building an SPD which aims to eliminate all error types bar that of undercoverage and then adjusts counts for undercoverage using Dual System Estimation (DSE) methods to obtain population estimates.
This paper explores the advantages and disadvantages of both methods before considering how they could be integrated to eliminate the disadvantages.
Many NSIs will be considering similar challenges when compiling annual Census like population estimates and this paper aims to contribute to that discussion.

1. Introduction

Statistical agencies in many countries are investigating methods for replacing traditional census based population estimation systems. Not every country has a Central Population Register (CPR) which can be easily used as the basis of directly compiled population statistics. Ireland is one such country. Central Statistics Office (CSO), Ireland, like many statistical agencies, has been investing significant resources into the exploitation of administrative data sources for statistical purposes (Dunne, 2015). As part of this effort CSO has been investigating new methods for the compilation of population estimates.

The first step in compiling population estimates is the compilation of a SPD from administrative data sources. The simple idea behind an SPD, is that it can be used instead of a CPR to count persons in the population for a given reference point or reference period.

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 can arise with respect to the 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.

Ireland, to date, has investigated 2 differing approaches to the compilation of population estimates from administrative data. The first, labeled in this paper as the *simple count method*, is based on building an SPD which minimises the overall number of individual record errors such that simple counts from the SPD will provide population estimates. The second, labeled in this paper as the *estimation method*, is based on building an SPD which aims to eliminate all error types bar that of undercoverage and then adjusts counts for undercoverage using Dual System Estimation (DSE) methods to obtain population estimates.

This paper explores the advantages and disadvantages of both methods before considering how they could be integrated to eliminate the disadvantages.

2. Methods

2.1. Simple count method

The simple count method was was used to compile population estimates for reference year 2020 . The population was estimated at 5.2 million †. An age by gender breakdown is provided in figure 1 where the simple count and estimation methods are compared for reference year 2020.

The method takes a signs of life approach to compiling the SPD. The rules underpinning the signs of life are chosen in an intuitive manner to target one record in the SPD for each person in the population. The approach relies on minimising the number of errors when counting records in the SPD to estimate the population. In practice there will be errors and errors in the resulting population will also cancel each other out to some extent.

Data sources with respect to universal child benefit payments, primary school enrolments, post-primary school enrolments, third level and further education enrolments, self employment,

†Published as a frontier release at `https://www.cso.ie/en/releasesandpublications/fp/fp-ipeads/irishpopulationestimatesfromadministrativedatasources2020/` accessed on 2nd June 2023

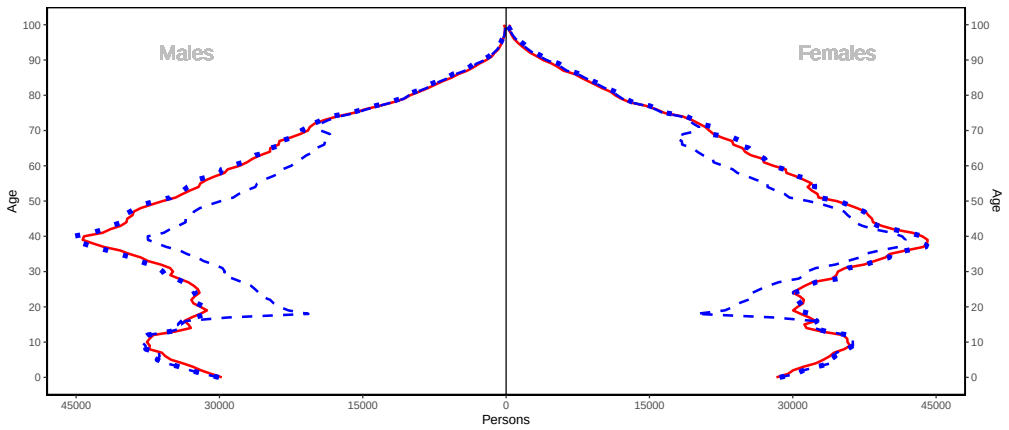

Fig. 1. A comparison of the simple count method with estimation method when used to compile population estimates by gender and single year of age, 2020. Red continuous line denotes estimates using *simple count method* - SPD compiled based on minimising number of errors. Blue dotted line denotes estimates using *estimation method* - adjusted counts from SPD compiled based on limiting type of errors to undercoverage - SPD also denoted using a blue dashed line.

employment, social welfare and pension payments were used to identify persons to be included in the SPD.

Location or place of residence was then assigned using a rules based approach to give more prominence to data sources that were considered of higher quality. Data sources used for assigning geography included rental registrations,local property tax for property owners and a listing of addresses maintained by the Department of Social Protection.

A limited number of attributes are also included in the SPD for each person. These include NACE sector for employees, nationality and whether a person is in receipt of a welfare payment along with core attributes such as age, gender and nationality.

There are some drawbacks. It is acknowledged in this approach than an adjustment to the rules for including a person in the SPD can impact directly on the population counts. However, if the rules are applied in a consistent manner from year to year and the underlying data sources are stable and robust in their operation, it can be argued that errors introduced are systematic and, therefore, shouldn't impact observations about the changing nature of the population from year to year. Another drawback is if there is a change in the nature of the data in the underlying data sources or if a data source becomes unavailable. Changes in underlying data source can come about due to a change in population behaviour with respect to interactions with the respective administration system, a change in rules (or implementation of rules) for the operation of an administration system or some other reason.

This approach has the advantage that it can compile coherent cross tabulations of estimates for the population by simply counting records for each table cell. This can be done for any attribute derived from the variables contained in the data sources contributing to the SPD. This method has only been applied for one reference year - 2020.

2.2. Estimation method

The estimation method comes from the Irish PECADO (Population Estimates Compiled from Administrative Data Only) project. More details including preliminary estimates for 2011 to 2016 can be found on the proceedings of the CSO Administrative Data Seminar, December 2018‡.

The estimation method when applied using selected administrative data sources for reference year 2020 estimated the population of Ireland at 5.3 million persons. An age by gender breakdown is provided in figure 1 where the simple count and estimation methods are compared for reference year 2020.

At its simplest, the estimation method is a 2 step process. In practice, an iterative process is used with an extended DSE toolkit§ to ensure the estimates are robust and can be defended from a methodological perspective. The toolkit can also be used to deal with suspect records and overcoverage errors in list A. Figure 2 illustrates the iterative process applied to ensure estimates are robust.

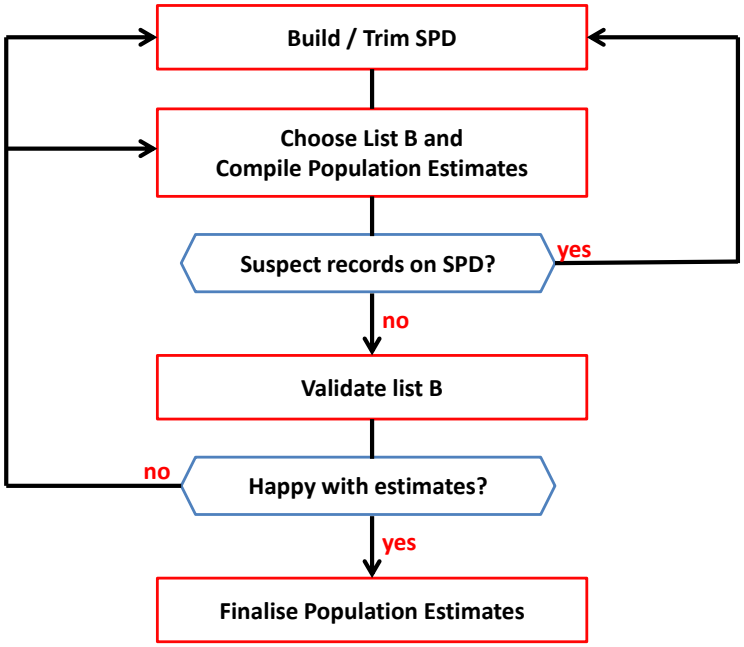

Fig. 2. High level process map for compilation of population estimates in the Irish PECADO project.

The first step involves the compilation of an SPD with only one type of error with respect to the population - undercoverage. However in this instance the SPD is compiled based on

‡https://www.cso.ie/en/csolatestnews/eventsconferencesseminars/administrativedataseminars/7thadministrativedataseminar/ accessed on 2nd June 2023

§Extended toolkit is documented in Dunne (2020)

applying a strict Signs of Life criteria over a similar set of data sources as that for the count method. The primary difference is that the criteria used for the SPD in the *estimation method* is to reduce the types of error to be dealt with down to one - that of undercoverage. The strict criteria have the purpose of ensuring that all records included in the SPD represent a person in the population, but the SPD does not necessarily contain a record for every person in the population. Undercoverage is expected.

Once the SPD is compiled, a designated data source deliberately excluded from the compilation of the SPD is used as a second list in a DSE setup to adjust SPD counts for undercoverage to obtain population estimates. The designated data source in this instance is a data source compiled from driver licence data and only includes records for persons who renewed their licence or applied for a new licence in the reference year. Post stratification by age, gender and nationality grouping is also used to enable disaggregation by these groupings and protect against any impact from heterogeneity in list B capture rates across these groupings. The tools/methods used to compile the population estimates are described as the PECADO toolkit.

An innovative aspect of this extended *PECADO* toolkit is that it revisits the DSE setup, in particular the assumptions used byWolter (1986), and restates the methods such that the assumptions are relaxed and restated as three primary assumptions with a fourth included to enable variance estimates. Denoting the 2 lists as list A and list B, the three primary assumptions are

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.

The additional assumption used to enable variance estimate compilation relates to *independent capture* of persons in the population on list B.

The *PECADO* toolkit also extends the traditional DSE methods such that parts of the SPD can be evaluated for erroneous records including overcoverage. This is an important extension as it now allows validation of the no erroneous records assumption when compiling estimates. In more general terms, this extension allows DSE methods to be used in the treatment of overcoverage errors.

The relaxation of the assumptions and the ability to be able to extend the methods themselves provides for a far broader application domain for DSE methods. One application that springs to mind is the replacement of the traditional post enumeration survey as part of the traditional Census with the use of a simple administrative list with the application of DSE methods.

Dunne (2020) contains a detailed presentation of the PECADO toolkit along with a presentation of estimates for reference years 2011 to 2016. Dunne (2020) also presents a reasonable argument for the robustness of the population estimates compiled.

The drawbacks to this approach is that there is no complete dataset for the population. The SPD compiled as part of this approach can contain significant undercoverage issues and as such it is not so easy to generate various cross tabulations for the population in a coherent manner.

3. Discussion and proposed combination of methods

In considering a comparison of the two methods, the simple count method with the estimation method, in figure 1 we see the two methods are broadly comparable.

The comparative strength of the estimation method are that it can be defended as a robust set of estimates from a methodological perspective while the comparative strength of the simple count method is that cross tabulations are simply derived by counting over the various dimensions in the SPD.

A consideration of the underlying methods and there comparative strengths leads to a proposal that combines both methods to leverage their comparative strengths. The proposal, in a simple form, can be described as follows:

First, compile an SPD from underlying data sources (holding a suitable data source back to use as list B in a DSE setup) that has also an attribute that scores each record on whether you consider the record to be *sure* (100% confident that it belongs to population) or *possible* ($<$ 100% confident record belongs to population but there is some probablity it does). For example, an SPD could be compiled with 1,100 records of which 900 are marked *sure* and 200 are marked *possible*

Second, compile benchmark population estimates using DSE methods where list A is the subset of the SPD where all records are marked as *sure* and list B relates to the data source that has been excluded from the compilation of the original SPD. In our example, population estimates could now be compiled with a suitable list B and list A containing 900 *sure* records to obtain a population estimate of 1,000.

Third, top up the records in list A to the benchmark population estimates using a probability based selection of records marked *possible* from the SPD. This creates a new SPD that can now be used for cross tabulations while summing to the population estimates that can be defended from a methodological perspective. In our example, list A could now be topped up by selecting from the 200 *possible* records in the SPD with a probability $0.5 = (1000-900)/(1100-900)$. In practice, some scoring system can be deployed to weight the probabilities of inclusion of *possible* records in the SPD.

Two challenges remain with both methods described here with respect to meeting demands for detailed population estimates; detailed geographical disaggregations and household composition. Address information on administrative data sources may be out of date or incoherent with other data sources and as such is not always accurate or up to date. It is difficult to deploy rules for assigning persons to detailed geographical location when the quality of address information on administrative data sources is varying, inconsistent and incoherent. Extending the toolkit to deal with domain misclassification as part of the *estimation method* contains challenges in dealing with small numbers associated with detailed geographical breakdowns. In general, only partial household relationships are captured on administrative data systems where it relates to a form of payment or relief in Ireland; for example a parent in receipt of child benefit payment will have a parent child relationship captured (note, even these administrative relationships may not mirror real world living arrangements - it could be the case that the child does not live with the designated parent and may in reality reside with another parent/guardian). The population consisting of third level students is a particularly difficult cohort to pin down in terms of geographical location and household composition, it is quite difficult to determine whether they are residing in some form of student accommodation or with their parents based on information provided in administrative systems.

In conclusion, the authors believe it is possible to compile population estimates from administrative data sources without the requirement of a public administration systems having a Central Population Register. The work undertaken to date shows this possibility, however, more work is required in developing the respective methods to address outstanding challenges, most notably to provide statistical detail on household composition and geography.

References

Dunne, J. (2015). The Irish Statistical System and the Emerging Census Opportunity. *Statistical Journal of the IAOS*, 31(3):391–400.

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

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