

Developing the Statistical Infrastructure for Frames and Population Datasets at CSO, Ireland

Sorcha O'Callaghan

Sorcha.OCallaghan@cso.ie

Tim Linehan

Tim.Linehan@cso.ie

Summary. In a previously published paper, [To count or to estimate: A note on compiling population estimates from administrative data](#), CSO Ireland discussed how it 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 Spine (SPS). A note on terminology at this point. CSO considers a statistical 'spine' to contain the ever-resident population, and annual 'usual resident' admin datasets are referred to as Statistical Population Frames (SPF). Ideally an SPF will have one record for each person in the population containing the relevant attributes. The ideal SPF then allows compilation of statistics by simply counting over records. Ireland, to date, has investigated two different approaches to the compilation of population estimates from administrative data. The first, was referred to as the simple count method and uses a rules based methodology to determine Signs of Life (SOL) from administrative datasets; and the second, was referred to as the estimation method and was based on using Dual System Estimation (DSE) methods to obtain population estimates. The previous paper explored the advantages and disadvantages of both methods.

Since this paper was published, CSO has decided to adopt the simple count method. The decision was based on good availability of a wide range of granular administrative data and a preference for explainability and hence transparency. This count-based method, called Irish Population Estimates from Administrative Data Sources (IPEADS) has been published for four annual iterations (covering reference periods 2020 to 2023) as a CSO Frontier (or in-development) statistic and now is currently under consideration for designation as an Official Statistic for individual statistics.

Furthermore, CSO is considering the broader statistical infrastructure required for using administrative data based on statistical spines. CSO is considering other statistical spines (and annual frames) including a household spine and a building spine. This

paper sets out at a high level the requirements CSO is considering in this regard along with anticipated challenges. Many NSIs will be considering similar challenges, and this paper aims to contribute to that discussion.

Abstract

Many countries do not maintain a Central Population Register (CPR) that can serve as a direct foundation for producing population statistics, and Ireland is among them. A necessary preliminary step in this process is the construction of a Statistical Population Spine (SPS) and annual Statistical Population Frames (SPF) derived from administrative records. The underlying principle of a Statistical Population Frame is that it can function as a substitute for a CPR by providing a count of individuals in the population at a specified reference point or over a defined reference period. An ideal SPF would contain one record for every statistical unit within the target population, with each unit uniquely identifiable through a stable identifier.

Although Ireland lacks a CPR, it does maintain a comprehensive system of unique personal identifiers, the Personal Public Service Number (PPSN), which is stored within a central records system (CRS). This CRS represents the ever-resident population i.e. any person ever assigned a PPSN. Ireland's SPS is compiled from administrative records in several domains including children, education, social welfare, tax, health and property ownership and rental. Signs of life rules are used to determine the usually resident population each year to generate the SPF. CSO enriches these SPFs with address information including Ireland's Eircode used as the linking variable for buildings/dwellings. The Eircode was first introduced in Ireland in 2015 and is state of the art. A key driver for the introduction of Eircodes was to deal with the challenges faced with non-unique addresses in Ireland. These include difficulties encountered by the emergency services in locating addresses due to 35% of all properties in the state sharing an address with other properties. This is the highest figure in the OECD and has a particularly negative impact for people and businesses in rural Ireland who require better access to goods and services. Over ten years later, it is used widely among the public, businesses and public sector.

CSO is now considering how these SPFs can be used as sampling frames for social surveys where the individual is surveyed. This will require consideration of how attributes can be added for post-stratification, weighting and calibration etc. Some of these attributes may have to come from the enumerated census and hence imputation strategies for persons not on the last census need to be considered. Some countries may have two separate administrative population registers for these two functions (population estimation and frame generation).

Furthermore, CSO will develop a building spine. This will cover all buildings in the state. Ireland does maintain a dataset of Eircodes, addresses and building information but

before this can be used as the basis for official statistics, quality assessment needs to be undertaken. Good metadata and methodological information are lacking on this dataset. We are aware that vacancy and building type are not considered accurate. Before CSO can use this dataset as the basis for statistics related to the stock of commercial and residential properties and for other building related statistics (climate and energy) area we need to establish a quality framework and consider augmentation with additional data.

An admin-based household spine is also under active consideration. This would allow timely indicators to be developed for pertinent issues such as fuel poverty and housing affordability. Many social survey frames are based on households and use the enumerated census as a basis. However, the census becomes out of date in the intercensal period. IPEADS will also publish demographic information at the household and family level.

Rules for assigning people to households are complex, as are common definitions for attributes such as household income. An agreed definition of household income derived from administrative sources is challenging, income as defined in the SILC survey may not match the definition of income used by the revenue collecting agency and those interested in developing an indicator around fuel poverty may have another view entirely. Ideally, we will need a set of principles that can be applied to administrative definitions to derive the required statistical concept.

While CSO envisages that the person and building spines will be compiled from administrative sources (with some census attributes added) it is possible that the enumerated census is the best starting point for the household spine with administrative data used to update it in the intervening years.

Concluding Remarks

CSO would like to start discussions with countries working in this area regardless of whether those countries have administrative registers or not. There are three main topics that we would like to open discussions on. The first topic is methodological guidance on using SPFs as sampling frames for social surveys covering topics such as required methodological checks, adding attributes from enumerated census or admin sources and imputation approaches. Secondly, we would welcome discussions on architecture and technical design, exploring how national statistical offices build scalable solutions, including statistical research services and the generation of national data assets for frequently asked research or policy questions. Thirdly, we would welcome collaboration on defining rules for constructing household spines/frames, including challenges related to definitions of households; how best to exploit or combine census and admin data to get best quality spines (timeliness and accuracy) and arriving at common definitions of household income.