Evolution of the Person Census and the Estimation of Population Counts in New Zealand, United Kingdom, Italy and Israel.

Antonella Bernardini [1] , James Brown[2] James Chipperfield [3], Christine Bycroft [4], Angela Chieppa [1], Nicoletta Cibella[1], Gary Dunnet [4], Michael F. Hawkes [5], Ahmad Hleihel [6], Eleanor C. Law[5], Daniel Ward[5], Li-Chun Zhang [7]

1 Italian National Institute of Statistics

2. University of Technology Sydney

3 Australian Bureau of Statistics, ABS House Belconnen, ACT, Australia M: 61 2 421086207

4 Statistics New Zealand

5 Office of National Statistics

6 Israel Central Bureau of Statistics

7 University of Southampton

Keywords:

Population Census; Administrative data; Statistical Registers; Coverage estimation

Abstract

A Census of a nation's people and housing provides statistics about its health, income and social structures at a local level. While the demand for these statistics is unchanged the way they are collected is changing in many nations because of common drivers: cost pressure, web-based collection, decreasing response rates, environmental shocks and the availability of administrative data. Within this context, this paper gives an overview of the evolution of the Census in Israel, Italy, New Zealand, and the United Kingdom and thereby provides an insight of the challenges and solutions of the modern Census.

1. Introduction

A Census of a nation's people and housing provides statistics about its health, income and social structures at a local level- a key statistic being the count of people, by age and sex. Governments and communities use these statistics in their funding and planning decisions. While these statistics could be provided at broad levels, by surveys, they would not be accurate enough to support local level decisions. There are an established set of countries that have long utilised register-based approaches to collate their decennial censuses, particularly the Nordic countries of Europe, but the ongoing demand for these statistics is challenging the notion of a traditional enumeration in many nations without that established history of (statistical) registers, because of common drivers.

The first driver is cost pressure. A traditional census involves a large workforce to distribute, collect and process Census questionnaires. This has been off-set to some extent by web-based data collection. There is also a significant cost in terms of response burden. Second, historically decreasing response rates are compromising the accuracy of census statistics. Third, environmental shocks such as weather events and the covid pandemic, affecting field operations, has led to delays of censuses that were planned during 2020 and 2021. Fourth, administrative data is available about a significant proportion of a nations' population at low cost. It often covers vulnerable populations, of particular interest to government, particularly well as they access government services.

These drivers are behind the modern evolution of the census in many countries. On this topic, we have reports from the national statistical organisations of New Zealand, United Kingdom, Italy and Israel. The common themes in these reports are interesting. First, while censuses a involve a long-term planning horizon, environmental shocks have presented opportunities for NSOs to speed up the evolution of the census, particular in their use of administrative data. Second, community trust is needed to support NSO use of administrative data. Third, administrative data are not collected for statistical purposes. An NSO may have little control on content and changes to processing of administrative data (e.g. family relationships are not well captured). Furthermore, they can be out-of-date, not include records for residents (under-coverage), and include erroneous records for non-residents (over-coverage). In the latter case, indirect indicators, or 'Signs of Life', are used to predict whether a record belongs to a resident. However, despite a trend to increase reliance on administrative data, the role of surveys remains the key to assuring quality in population counts.

Statistics New Zealand (Stats NZ) (Section 2) and the Office for National Statistics (ONS) (Section 3) are drawing on considerable research work to determine the exact role that administrative data will play in their upcoming censuses. The ONS is considering operating without a Census. In preparation for such a scenario, this paper describes their research into using administrative data to count their population over time. Stats NZ give an interesting historical perspective of their census and demonstrate the potential of using administrative data to improve the quality of a traditional census. The 2011 Italy Census (Section 4) was the last traditional Census. Since that time they have used administrative data as the basis for counting their population, using surveys to measure its over-coverage and under-coverage. For some time, the Israel Census (Section 5) has been based around its population register. They discuss the quality issues with the register and mention the surveys and other methods to overcome them. Section 6 suggests two alternative frameworks for these Census models.

2. Person Census in New Zealand

2.1 A Historical Review

The Origins

The first population census for the colony of New Zealand was taken in 1851, but only included European settlers [1]. The 1858 Census was the first attempt to collect comprehensive statistics on the indigenous Māori population. The census has been held every five years since 1881, with very few exceptions (the census in 1931 was abandoned due to the economic depression, and in 1941 due to the second world war, and most recently an earthquake caused a postponement of the 2011 Census). Until 1951 when they were integrated, the Māori and European populations were enumerated in separate censuses using different schedules and collection methodology. These censuses followed the same basic traditional approach of making contact with all households to obtain completed questionnaires.

In response to new technology, an online questionnaire was first trialled in the 2006 Census, but paper forms were still by far the dominant mode. A traditional full field enumeration census is an extraordinary undertaking by any measure. The two most recent censuses have faced unprecedented and unforeseen challenges. While these situations were difficult at the time, they have been catalysts for change in the way the New Zealand census has been undertaken.

The 2011 Census becomes the 2013 Census

On 22nd Feb 2011, 2 weeks before census day and with field work in full swing, a major earthquake struck the Canterbury region, with its epicentre in the city of Christchurch. Christchurch was also the centre of census operations. It became obvious very quickly that the census could not go ahead, due not only to the scale of destruction in Christchurch, but because of its impact on census staff. The census was rescheduled to March 2013, two years out, which was the minimum period needed to set up the census operation again. There were, at the time, no alternative ways to obtain census information other than asking all households to complete census forms, and the two-year delay offered no room for innovation beyond what had been planned for 2011.

Census transformation strategy in 2012

Partly as a response to this situation, a census transformation strategy was developed and agreed by government in 2012 [2]. The strategy has two parallel strands. The first strand was a short- to medium-term focus on modernising the current census model and making it more efficient, with the first implementation in the 2018 Census. The second strand had a longer-term focus exploring the feasibility of a census based largely on administrative data and supported by sample surveys.

The longer-term census transformation investigations are focussed on assessing the quality of administrative sources, considering how that quality might be improved and what transformations are necessary to provide fit for purpose statistical information. Research has been undertaken using the linked administrative data in Stats NZ's Integrated Data Infrastructure (IDI) [3]. The IDI includes a wide range of government data, Stats NZ's household surveys and the 2013 and 2018 Censuses. Because there is no national population identifier in New Zealand, linkage in the IDI is mainly undertaken using probabilistic methods [4].

The body of knowledge has grown substantially over time, to the point where we now have a good understanding of the census variables where we have high quality administrative data sources, and those for which a survey will still be needed [5]. An essential feature of the long-term use of administrative data for census information has been the development of an *admin New Zealand resident population* derived from linked administrative data [6]. While linkage errors do exist, we have found they have little impact on our ability to produce high quality data for census.

The 2018 Census

The two strands of the strategy are interdependent, and the 2018 Census was aligned with the longer-term vision by including a goal to increase the use of administrative data. The main use of administrative data planned for 2018 was to construct an address frame from administrative sources, which has developed some of the organisational infrastructure required for long-term change. The 2018 Census also planned to use alternative sources to improve data quality in the context of missing data for census questions, based on the findings of the administrative data research.

Another key aspect of the modernisation involved moving from a largely paper-based census, with online as a supplementary mode in 2013, to a digital first approach in 2018, with paper forms as an alternative mode as needed. In this, New Zealand was following a model developed by Statistics Canada and successfully implemented in their 2011 and 2016 censuses.

The value of the joint census transformation strategy became clear in the light of lower than anticipated response rates in 2018, which have considerably extended the role that administrative data played in the 2018 Census. While some aspects of the 2018 Census modernisation were successful, we faced major challenges when implementing the new collection model, and the overall level of response was lower than expected. While there was a systematic undercount across the whole population, it was clear that non-response was more concentrated in some geographic areas, and that groups who are typically harder to count in the census were disproportionately affected in 2018.

Previous censuses have included an adjustment for non-response, through inclusion of 'substitute' records, a form of unit imputation, and a similar approach was planned for 2018. Once these collection issues became clear in mid-2018 it was found that the previous unit imputation approach was likely to give biased results given the extent and nature of the non-response. Stats NZ then embarked on development of new methods using administrative records to count people who were missed by the census field collection, replacing the use of 'substitute' imputed records in previous censuses. The new methods were based on the derived admin New Zealand resident population already developed through the research undertaken by the longer-term census transformation programme.

This process relied on the good understanding of the strengths and limitations of the administrative data quality that we had gained from previous research. Statistical methodologies had to be developed to account for the known quality limitations. These included aspects such as the removal of over-coverage from the admin resident population, measurement and adjustment for linkage error, and determining the probability that administrative address information would place individuals in the correct household or correct small area. This work was not planned as part of the build-up to the 2018 Census, and the new methods were developed under considerable time pressure to minimise delays in releasing census outputs.

The 2018 Census is effectively no longer simply a traditional full-field enumeration census, but a combined census model, where the census dataset is made up of census forms received through field collection, supplemented by high-quality administrative records for people from whom we have not received a response [7]. We called these records 'admin enumerations'. Characteristics of these people and dwellings are also obtained from linked administrative data (and the previous census in 2013), where possible [8].

The final 2018 Census dataset consists of 89 percent census responses and 11 percent admin enumerations. Admin enumerations make up almost a quarter of the Māori and Pacific populations, compared with 10 percent for the European ethnic group, an indication of lower participation for people of Māori ethnicity and the Pacific ethnic group in the 2018 Census.

It is clear from our results that the admin resident population does include many people who are typically hard to count through census field collection. The use of administrative sources for their characteristics means that the census data includes real data about real people and gives a higher quality result than the use of imputation or leaving missing values. In previous censuses, only age and sex were imputed for substitute records and other characteristics were 'not stated'.

The value of the census questionnaire component is evident for census variables where no alternative sources were available. These variables consequently have higher levels of missing data in 2018, and those sub-groups with lower response rates to the field collection are more adversely affected.

The change in methods has resulted in a disruption to time series which can be difficult for customers. In some cases, the 2018 Census data is better than previous censuses, for example the important population counts by ethnicity are more complete in 2018, because administrative values can be used instead of leaving missing data. However, for variables where there are no alternative sources, the 2018 Census is more impacted by any biases due to missing data than previous censuses.

2.2 Looking forward

The next census in 2023

The use of administrative data in the 2018 Census has resulted in a fundamental shift from the traditional census model previously used in New Zealand. Administrative data is now purposefully included in the next census in 2023 which has a 'combined census' model by design [9]. However administrative data is still very much a complementary data source used to fill gaps in survey responses, and the success of the 2023 Census depends on achieving high response rates to the field enumeration.

Planning this use of administrative data from the outset allows for a more measured approach for the 2023 Census. It gives time for statistical methods to be reviewed, but more importantly allows time for development in partnership with Māori, and for wider engagement with customers, which was not possible in the lead up to 2018 Census releases due to the time pressures we faced.

Continuing development of administrative data for census information

The potential for use of administrative data for census information continues to be explored. We are now moving beyond a purely research phase to demonstrating the potential through an experimental *administrative population census* (APC) [10]. The experimental APC has been released to show what can be achieved currently with administrative data, and to provide customers the opportunity to be involved in further development.

The APC builds on experimental population estimates from linked administrative data [11] released in 2016, 2017, and 2018, and extensive and detailed research on the quality of administrative data for census variables, but takes this work one step further. The first release in August 2021 is an annual time series from 2006 to 2020, compiled from underlying longitudinal unit record data. It provides estimates of the New Zealand resident population and demographic and identity variables: age, sex, geography, ethnicity, Māori descent, birthplace and years since arrival in New Zealand [12].

Results for this first version of the APC in 2021 show good agreement with the expected population patterns and distributions of census variables at high levels of aggregation. There is a small net undercount of the total population compared with official population estimates, although the APC population is always larger than the corresponding census counts.

Further versions of the APC will be released in 2022 and 2023. They will take the time series forward and include variables for topics such as work, income, and qualifications.

Advantages over a traditional census are that results can be updated annually, and with low effort once routines are in place. The underlying unit-record data is inherently longitudinal which offers much richer analysis potential than the traditional cross-sectional census dataset.

Censuses after 2023

The original drivers for the census transformation strategy in 2012 remain relevant today. In 2012 there were concerns about the rising cost of the traditional model due to the growing population and increasing difficulty of obtaining responses, and whether this would be sustainable in the future. These concerns were borne out in 2018 and remain a risk for the 2023 Census. They were also seen to be opportunities from the increasing availability of administrative data. Since 2012 the capacity to use government administrative sources for statistical purposes has grown considerably, much of this supported by the IDI which now has up to a hundred new social research projects each year.

Legislation requires the next census to be held in 2028. The 2018 and 2023 Censuses are stepping-stones on a pathway of increasing use of administrative data and it is already clear that administrative data will play a valuable role in future censuses. However, no decisions have been made on the approach to census beyond 2023. Early consideration is now being given to potential options which will need to be firmed up by mid-2023 for the next decision point. Good progress has been made on some of the requirements for shifting to a census primarily based on administrative data, but considerable work remains.

The administrative data provide a strong foundation for core census demographic counts, and for around half the characteristics of individuals. The current lack of suitable alternative sources for iwi affiliation (Māori tribes) is a critical gap which is likely to take some time to remedy.

It is also clear that administrative data cannot meet the full range of topics traditionally collected by the New Zealand census. Administrative data for household and family information is not yet accurate enough, and alternative sources for variables such as language spoken and activity limitations are very limited. Under an 'administrative-first' approach, a large-scale sample survey will be needed for variables such as these. Indicative design work suggests this would be a five percent sample of the population each year and would provide annual updates of census information.

The sample is designed to provide good estimates down to Statistical Area 2 (a small area geography of around 2,000 people). However, there will be a loss of highly detailed information in comparison with the current census that will impact smaller population sub-groups. Loss of data for small groups and at low levels of geography is an enduring concern for customers.

We also need to consider how this shift in the workload pattern for census from once every five years to an annual sample will impact on other household surveys, and Stats NZ's ability to manage this higher level of surveying.

The IDI has shown that we do have the ability to link data sources, despite the lack of a national personal identifier that is common in countries that have already shifted to a 'register-based' census. However, the IDI is designed as a research environment, not for statistical production. Work is currently underway to develop an integrated data environment as an enterprise asset based on the ideas of a register-based statistical system. This new data infrastructure is seen as necessary before a primarily administrative-based census could be efficiently implemented in a production process.

New legislation under the Data and Statistics Bill (once passed) provides the legislative environment for different approaches to census taking. It removes the highly prescriptive approach to the census of the 1975 Statistics Act while retaining the five-yearly cycle. Other provisions clarify and enhance the ability of Stats NZ to obtain access to administrative and other data sources for official statistics and research purposes.

Our experience in the 2018 Census highlighted the lack of progress on understanding public acceptability ('social license') and building trust and confidence in Stats NZ's use of administrative sources for statistical purposes. To mitigate this, in the lead-up to the 2023 Census, we have been engaging with our customers and partners, and with tools such as the APC, to build their understanding and confidence in the use of administrative data in the combined census model. In addition, we have been working closely with the NZ Privacy Commissioner and regularly undertaking attitudinal surveys as a barometer of New Zealander's appetite and understanding of using administrative data for statistical purposes. In all of our interactions, we aim to be open and transparent about our use of administrative data.

In making a decision on the future of census, the Government Statistician will take wider considerations into account, including the ability to meet information needs under a different approach to census, having appropriate legislation in place, and retaining trust and confidence in Stats NZ. Even with favourable conditions, a census primarily based on administrative data with sample surveys in a supporting role would be a paradigm shift. We should not under-estimate the impact of such a major change on Stats NZ's people, the new capabilities required and new governance and organisational structures.

2.3 Reflections

Ideally changes to the census model are planned and tested well in advance, and able to be implemented with confidence. This is more viable when change is incremental. However, as we have seen in the New Zealand 2011 Census, change can be forced by circumstances outside the control of the agency, and in 2018, major planned changes are not always successful.

After the earthquake in 2011, there was no option but to stop the census mid-operation and run it afresh two years later. The difference in our ability to recover from the 2018 Census was due to two factors: the contingency capacity we had available, and processes that were put in place to manage uncertainty. A key contingency was having the long-term pathway towards greater use of administrative data already established and partly tested. The linked data environment that provided access to cross government data was also already in place and the legislative compliance established. Stats NZ also had experienced, expert methodological capability who were focussed on developing solutions for the 2018 Census. There was still considerable uncertainty as we worked on new methods. Two key mechanisms helped manage this uncertainty. An external panel of experts experienced in census data was set up and provided an excellent sounding board, offering advice and challenges to our thinking. Commitments for publication date were kept loose until we had a firm idea of the solutions.

Clearly, it is much better to plan changes rather than being reactive. A successful field collection in 2023 will be critical in determining the pace of change. In terms of shifting to an 'administrative first' combined model, New Zealand is well placed in terms of the suitability of administrative data sources, and our ability to apply appropriate statistical methodologies. But this is only part of the larger picture. It is important to have a clear end goal. The challenge for Stats NZ for future censuses after 2023 is to find the correct balance between ambition and what is achievable in practice, and to be able to take our Māori partners, customers, and the general public with us

3. Population Counts using Administrative Data instead of a Census: United Kingdom

3.1 Context

In 2023 the Office for National Statistics (ONS) will make a recommendation regarding the future of the population and migration statistics system in England and Wales. This recommendation will consider the need for another Census in 2031 and explore possible designs for a transformed statistics system that can meet user needs using a range of existing and new data in a post-Census world [13].

ONS currently produces population size and characteristics statistics for England and Wales from a decennial Census and post-Census coverage survey using dual system estimation [14]. Equivalent statistics for Scotland and Northern Ireland are produced by National Records of Scotland (NRS) and the Northern Ireland Statistics and Research Agency (NISRA), respectively. Intercensal population statistics are produced annually at mid-year by rolling forward Census estimates using a cohort component method and a range of administrative and survey data [15]. These mid-year estimates contain national and sub-national statistics by single year of age and sex. Error in the mid-year estimates increases year-on-year after each Census until they are rebased using the results of the subsequent Census. ONS does not currently produce additional intercensal population characteristics statistics, such as small area ethnicity by age-sex group.

Transforming the population statistics system in England and Wales to operate without a Census poses several challenges, not least of which is the lack of an authoritative UK population register with persistent identifiers for individuals across all government-held administrative data. In the absence of a population register, ONS is exploring options for administrative-based population estimates (ABPEs) produced from integrated administrative data combined with a population and characteristics coverage survey [13, 16].

One method that ONS is exploring to support the production of multivariate population statistics from integrated administrative data is *fractional counting* [17]. Fractional counting is a model-based approach to fractionally weight record-level integrated administrative data. The models attempt to predict the probability that each component of the data describes a real individual in a target population, thus reducing elements of over-coverage and resolving attribute value conflicts across administrative sources. Other work is exploring rules-based inclusion and edit methods for the same purpose [18].

The fractional counting models could result in more robust population estimates than rules-based integer counting because it more accurately captures uncertainty in the underlying administrative data and its linkage. Consider a situation where two administrative sources record an individual living in two different locations, and both are equally likely to be correct. A deterministic rule to resolve this conflict would allocate the individual fully to one location. A fractional counter, however, would allocate the individual fractionally to both locations in proportion to the predicted probability of each being correct. Counting fractionally could therefore reduce the bias of ABPEs. This is particularly advantageous when aggregating population counts across multiple characteristic attributes as in regular Census population characteristics outputs.

While fractional counting has the potential to address some challenges when producing ABPEs, particularly over-coverage and item value conflicts, it will not be a panacea. Fractional counting would form one methodological component of a statistical pipeline that would also include other methods including data linkage, estimation and adjustment, and statistical disclosure control. These additional methods are critical to the production of ABPEs, and their quality will in large part determine the quality and robustness of ABPEs with or without fractional counting

Here we present our initial research using supervised learning to build a fractional counter for the provision of multivariate population statistics in England and Wales. There are four main components to this work: (1) construction of an integrated population dataset from which to produce population estimates, (2) creation of models to fractionally weight integrated data, (3) development of methods to update the models over time, and (4) a process for auditing these models at regular intervals.

3.2 Building an extended population dataset

To produce ABPEs we require an integrated population dataset (IPD) constructed through the linkage of several administrative sources. The IPD used here is constructed using anonymised historical data from the NHS Patient Register, Higher Education Statistical Authority (HESA), English School Census, and Welsh School Census.

An IPD constructed from linked administrative data alone is expected to describe a population less accurately than a population register augmented with administrative data. An admin-based IPD can contain several types of error. These include over- and under-coverage of individuals through erroneous duplication, inclusion, or exclusion; missingness within demographic and other population characteristic attributes; and domain classification errors where administrative data holds incorrect information, potentially in conflict across multiple sources. Any of these errors can result in biases, especially if they are structured across sub-populations. Errors in an IPD may be introduced by erroneous data linkage, operational errors within each administrative source resulting in incorrect data being stored, or lags during the collection and processing of administrative data resulting in stored data being inaccurate for a given reference date. In addition to error, definitional differences between administrative sources may result in genuine and potentially irresolvable conflicts over certain attributes, for example the use of different categorical frameworks to record ethnicity. Such genuine conflicts may be resolvable where definitions can be mapped across sources. Similarly, administrative conflicts over attributes that can change over time (like residential address or other geospatial attributes) may be the result of legitimate individual behaviour being partially captured by different sources. For example, when someone splits their time between two addresses they may be recorded at the first address in one source and the second in another. For statistical purposes we would usually try to classify one of these as a usual residence, but fractional counting would potentially allow us to capture such artefacts of individual behaviour in alternative population bases.

We assume that the IPD contains over-coverage but little under-coverage for the purposes of fractional counting. Other research has shown this assumption may not hold for currently available sources of administrative data and linkage methodologies [18], but we expect current levels of under-coverage to decrease over time as administrative sources are improved and new sources are acquired. Methods to deal with persistent under-coverage, such as adjustment via a coverage survey, will be addressed in future work.

When constructing an IPD the typical aim is to produce a single view of the integrated data that is as accurate as possible. In practice this means the IPD should include a single coherent entry for every member of the target population recorded in administrative data. When fractional counting, we are instead interested in predicting the probability that each possible version of a person described in administrative data exists in our target population. Therefore, we want to retain alternative attribute values for individuals as well as individuals outside of the target population in an *extended* population dataset (EPD). We can conceptually make a distinction between *real* individuals and *administrative* individuals, where administrative individuals are possible versions of real individuals as recorded in administrative data (potentially with multiple administrative individuals relating to each real individual).

3.3 Building a fractional counter from the extended population dataset

Our goal is to build and test a fractional counter that can assign a weight to each administrative individual in an EPD so that each can contribute to population estimates in proportion to their probability of accurately describing a real individual in the target population. This weight should capture the probability of inclusion in our target population (usual residents of England and Wales), the probability of correct placement within alternative statistical geographies, and the probability of correct description across alternative demographic and other characteristic attributes. To date we have tested a range of supervised learning algorithms in a two-stage fractional counter that predicts probabilities for target population inclusion and geographic placement.

The first stage of the fractional counter predicts the probability of inclusion in our target population framed as a binary classification problem (included vs. excluded) for every real individual observed in the administrative data; every administrative individual corresponding to the same real individual is assigned the same inclusion weight. The second stage predicts the probability of geographical placement across alternative possible locations framed as multiple one vs. rest classification problems (one for each possible location; correctly placed vs. misplaced).

Currently, we choose a single set of demographic and other characteristic values for every individual according to their most recent attribute values in the EPD. We are exploring options for extending the fractional counter to include population characteristics in one or more additional stages of modelling. A key challenge is poor coverage of some characteristics in administrative data [19]. We are considering various imputation methods to help solve this problem, for example fractional imputation [20].

We extract the model-predicted probabilities from both stages and their product yields the final fractional weight for each administrative individual. Within each stage the total probability across all alternative options should sum to 1, though we have not yet added this constraint to all our models.

We train our models on an EPD referenced to mid-year 2011 and linked to Census 2011. Individuals present on the Census and the EPD are flagged as members of our target population, and all other individuals are flagged as non-members. These inclusion flags are used as labels for training the stage 1 models. We do not yet correctly label members of our target population who are present in the EPD but did not respond to the Census, a type of Census coverage error estimated by the Census Coverage Survey. Individual placement in the EPD (currently by postcode) is flagged as correct or incorrect based on comparison between the administrative location and the Census location. We assume that the Census location is correct, and that the correct location is one of those available in the EPD (though we are currently investigating the implications of dropping the latter assumption). Both assumptions can be violated due to either the difference in reference dates between Census and the EPD, or error in the administrative data. These placement flags are used as labels for training the stage 2 models.

We have built alternative fractional counters using four supervised learning algorithms for classification: logistic regression, support vector machines, random forest, and gradient boosted trees. We will not explore the detail of model training and assessment here, but initial results suggest that all can approximate a correctly weighted population sample to varying degrees without respect to placement, showing qualitatively similar patterns of over- and under-coverage across sex and single year of age. *Correctly weighted* in this context means where target population membership and correct addresses are known and can be weighted correctly to 1 or 0 in the test sample. Current work is exploring how stage 2 weights influence placement in sub-national population estimates compared to rules-based integer methods.

3.4 Future work

Machine learning models are said to drift when they become less accurate over time. This occurs when the relationship between the dependent and independent variables changes but the modelled function describing this relationship is not updated; the real-world drifts while the model is fixed in place. In the context of fractional counting, such change could include fluctuations in coverage patterns across administrative sources, improving or weakening data assurance processes, and changes in data linkage quality. A method is required then to update the models in a fractional counter to prevent significant shifts in bias over time.

We are exploring how we might use an annual coverage survey to provide new training data to allow our models to be updated in an incremental learning regime. A population and characteristics survey linked to the EPD, in combination with other components of administrative data (e.g. death registrations and emigration status), could provide positive and negative labels to update our stage 1 and stage 2 models. We will also consider whether the design of this coverage survey could support some element of active learning, where the survey would target individuals whose inclusion would most improve the performance of the fractional counter.

We will also consider whether a fractionally counted EPD could be combined with a coverage survey to produce robust local area estimates, and whether the use of fractional counts rather than integer counts would alter the statistical properties of the estimates. Fractional counting is in many ways a weighting strategy, and while our statistics have long employed weighted data, we will consider whether fractional counting would require any changes in how users understand or use our statistics.

In the context of producing ABPEs with temporally-stable error we will investigate methods for estimating the uncertainty of fractional estimates alongside similar work underway to estimate uncertainty for integer estimates [21]. Additionally, we will assess methods for intermittently auditing the fractional counter and the accuracy of its outputs, and consider the need for a separate data collection exercise for auditing or whether it can occur as part of the model updating process.

3.5 Summary

ONS is preparing a recommendation regarding the future design of the population and migration statistics system in England and Wales. Part of this recommendation will consider the feasibility of using fractional counting in the provision of robust multivariate population statistics without a Census. The research presented here documents the early stages of this research, with the ultimate goal to produce a body of evidence regarding the feasibility of fractional counting. We will continue to publish updates as the work progresses.

4. The use of Administrative Data in the production of Italian Permanent Population Census Estimates

4.1. The Italian Population and Housing Permanent Census

Until 2011, Italian Population Censuses aimed to count all residents every ten years, with a complete enumeration of all units in target population. The censuses heavily relied on field-collection and required ISTAT (Italian National Institute of Statistics) to interact with local authorities. The 2011 Census was the last to make use of complete enumeration while introducing administrative records to support some field operations.

In 2016, ISTAT adopted a modernization programme involving statistical registers, to manage the statistical integration of administrative and survey data. The aim of this programme is to put the statistical registers at the core of the statistical production and to carry out 'registers-based' surveys to improve registers' quality, reducing surveys costs (largely by reduced field work) and response burden on citizens and businesses.

In line with this new framework, the Permanent Population and Housing Census (PPHC) strategy was adopted in 2018. This strategy aimed at integrating administrative data and sample survey data to obtain the annual counts of the usual resident population as well as other thematic Census outputs [16]. The register at the core of the Permanent Census is the Population Base Register (PBR), whose main administrative sources are the Local Population Registers. The PBR contains under and over-coverage of residents; moreover, to produce thematic Census outputs there is a need for variables not included in ISTAT registers. Two sample surveys, an Area-based survey and a List survey, have been designed to estimate the PBR coverage errors and gather other data needed to estimate all Census results. Both surveys are conducted annually in self-representative municipalities (i.e. with > 17,800 inhabitants), and once every 4 years in non-self-representative municipalities. At the end of the first cycle (2018-2021) all Italian municipalities will be sampled at least once. Out of a total 7,914 municipalities, about 2,850 are surveyed every year, for a total of about 1,500,000 households (of which 450,000 for the Area-based survey and 950,000 for the List survey).

The Area-based survey is built on a sample of addresses and/or enumeration areas (depending on the quality of addresses in a given municipality). The sample is drawn from an address register, rather than being drawn directly from the PBR, with the aim to count (interview by CAPI) every usually resident household, independently (or “blind to”) the PBR.

Conversely, the List survey is a sample of households drawn directly from the PBR. Interviews are conducted with a mixed mode technique (CAWI, CAPI, CATI), with a first phase of so-called “spontaneous response”, and a second phase of field follow-up of non-respondents.

4.2. Estimating Population Counts for the first waves of Permanent Census (2018-2019)

Resident counts by municipality are the primary Census output in Italy. According to PPHC design, Census counts are calculated by adjusting the PBR data by estimates of:

- over-coverage: inclusion in the register of individuals who are no longer in the municipality
- under-coverage: non-inclusion in the register of individuals who are on the municipality.

The capture-recapture model is adopted for direct estimates of the coverage errors. In a traditional census a Post Enumeration Survey (PES) measures the census undercount, with the census being the first ‘capture’ and the PES the second ‘capture’ [17]. Here, the PBR is the first capture, the Area-based sample is a traditional second capture for under-coverage while the List sample is a second capture targeting over-coverage [18, 19].

Indirect estimation with Small Area Models is used both to enhance the quality of direct estimates for sampled municipalities and to calculate estimates for non-sampled municipalities.

During the data processing steps of the first two waves 2018-2019, the theoretical design was partly modified due to fieldwork quality issues and administrative data were used to support the estimation of PBR coverage errors.

The Integrated Archive of Usual Resident Population (AIDA) was set up in 2015 to exploit all administrative sources useful for population estimation at the person level (e.g. Labour and Education archive, Tax Returns archive, Earnings, Retired, and Non-Pension Benefits archive, Permits to Stay archive). Administrative data are classified according to duration patterns, reliability of the administrative source and the association with other individual records (household relations for example) with the aim of predicting whether a record on the PBR belongs to a usual residence. These derived data are often referred to as *Signs of Life (SoL)*.

Non-respondents to the survey are classified as a usual resident if they show strong SoL, classified as activity in at least eight of the previous 12 months, in the same municipality where they are recorded in the PBR. Otherwise, non-respondents are classified as over-coverage under the hypothesis that someone not counted in the survey without a strong presence in the area is not a usual resident. The ‘raw' estimate of the over coverage rate in each municipality is given by the ratio of the number of sample units in PBR classified as ‘untraceable’ to the total number of sample units. Without using the signs of life, the 'raw' estimate of the over-coverage rate was 6.10% (at national level) instead of the adjusted estimate of 2.73%.

At the end of the process, a 'weight' is attached to each individual in the PBR, which 'corrects' for the coverage errors. Therefore, if over-coverage and under coverage are equal, the weight is equal to 1. In other words, each individual represents him/herself and therefore will be ‘counted’ as 1 in order to get the population count at the municipality level. On the other hand, if the estimated under-coverage of PBR is greater than the estimated over-coverage, the weight will be greater than 1.

4.3. The use of administrative data: investigation areas and estimation for 2020 counts

There are two main areas of further investigation. The first topic covers the combined use of survey and administrative data, making the most efficient use of administrative data integrated in the AIDA database and of the information from census surveys (List and Area). It is necessary to improve the accuracy of municipality level estimates, particularly for small municipalities, through the availability of multiple lists that enumerate the population of interest [20]. Supervised and unsupervised models (e.g. classification trees, latent class models) at individual level are currently being evaluated. Early results show clearly different patterns useful for estimation, identifying aggregates with different response characteristics: coverage rates are higher in large towns, lower for students, higher for retired persons (always above 90%) who make up 23% of Italians. By contrast, about 2% of Italians have no sign of life and foreign workers living alone in suburban areas are not well captured by the PBR.

A second topic for investigations is to use only administrative data for estimation of Census results. The 2020 COVID Pandemic was a strong pushing factor in this area. A deterministic approach to produce census counts on the basis of solely SoL is the solution adopted for the production of census counts in absence of surveys (due to the pandemic circumstances). We are evaluating results on calculating PBR coverage for year 2020 exclusively using deterministic rules based on SoL; these studies are very relevant also in the future perspectives for the Census round after 2021, for which we have a serious reduction in census budget for surveys.

A tentative evaluation of the quality of PBR performed exclusively by means of AIDA for the year 2017-2018 is given in Table 1. The Resident population on the basis of AIDA is equal to 59.714.128 individuals while in PBR it is equal to 60.529.243. More than 59 million are correctly reported according to strong SoL information (more than 98.2% of the total PRB). From year 2020, new administrative variables (e.g. Citizenship Income and household energy consumption) can help minimise coverage errors on the PRB.

4.4 Designing new solutions and future perspectives

The PPHC aimed at integrating registers, administrative data and sample surveys to obtain the counts of the usual resident population at a reference date, including the estimation of the coverage errors of the PBR and to produce thematic Census outputs. The data of the first two waves of Census surveys constitute a valuable source for studying how to integrate administrative data in the estimation processes. The best solution requires not only methodological aspects and testing model assumptions, but accounting for surveys' costs and designing and managing databases workflows. Building a new solution for improve quality and rethink the estimation process requires an interdisciplinary team; local authorities (public stakeholders) are to be considered too.

To extract knowledge from data is extremely important for designing a new solution, considering the complexity of the phenomenon and that a massive use of administrative data for population statistics is an innovative solution. The overall objective of this pioneering process is to find the best 'learner' for classifying signs of life and presence in PBR and predict usual residence in Italy at individual level.

In this sense, some steps accomplished to integrate survey and administrative data [21,22] and to produce a valuable framework for investigations are:

- pre-processing survey data to eliminate biased data:
- standardized definitions of SoL across administrative data, including the strength and duration of SoL;
- identify patterns of SoL where prediction of usual place of residence is uncertain.

During the pioneering process for the year 2020 using administrative sources we completed:

- construction and introduction of new variables (individual, family, travel routes, typology of municipalities) to classify people as residents;
- formalization of the residence criteria by means of a "grid map" for the classification of individual records on the basis of the variables identified in the previous points;
- deterministic rules using SoL for classifying individuals in BPR as "correctly placed", "candidates for under coverage" and "candidates for over coverage";

- use the above classifications to calculate and analyse in order to improve the estimates of the over and under coverage of the BPR.

Moreover, we have developed an experimental integrated database that can be easily used for analysis purposes and for model selection. First analyses allow the detection of some critical subpopulations (*grey areas* such as one-person households and persons without administrative signs like foreigners) that need further investigations and possibly changes for improving the quality of the surveys. Various estimation models are currently being evaluated.

5. Israel Censuses

5.1 - A Historical Overview:

Israel's first census was held in November 1948, just six months after the country declared independence. The major goal of the census was to register residents and establish a population registry. Because the census was conducted during a war, the results were incomplete, with the surviving Arab communities in the new state's territory accounting for the majority of those counted. Within 2-3 years after the census, the unregistered population was registered.

In Israel, the first census served as the foundation for population statistics. The Israeli Central Bureau of Statistics (ICBS) received monthly summary tables of births, deaths, address changes, and data on departures and arrivals in Israel from the Population Registry. Prior to 1961, population estimates were calculated using the present population (de facto).

Traditional censuses were conducted in Israel in 1961, 1972, 1983, and 1995. The population in all of these censuses referred to people who live in Israel on a regular basis (de jure). Between the censuses, from 1961 to now, population estimates were based on the findings of the previous census and were updated using civil registration as recorded in the population register from the date of the previous census to the date of the estimate.

Following the 1995 census, it was decided to increase the use of administrative data to estimate population in the census. This choice was made due to the difficulty of budgeting as well as a desire to lessen the response burden. The decision to go to a future administrative census was based on this conclusion. As a result, it was agreed that the 1995 census would be the last traditional census, and a variety of measures to examine the population registration were taken; monitoring population estimates at the person level and performing a sample census is one of them. The primary goal of the sample was to identify and repair the population registry's "errors." [28].

An integrated census was conducted in 2008, combining data from the population registration with data collected on the field. Each of the 3,000 statistical areas (population estimate target) is divided into tiny cells, with each cell containing an average of 50 households. A sample of around 17% of the cells was taken. Two lists of persons were obtained for each sampled cell. The first is a list of persons who are registered in the cell's population registration, while the second is a list of people who were discovered in the cell area during a "door to door" census of the region inside the cell limits. Using the dual system estimator, the population estimates were calculated by comparing the two lists. The Israeli population registered in the Population Registry was included in this census, but the foreign population not registered in the Population Registry was excluded. This is due to the lack of an administrative sampling framework for foreign arrivals in Israel [29], [30].

The 2008 census was eventually not used to transition to an administrative census because a rolling Census on a 10-year cycle was planned shortly after the census ended. In 2012, the first round of the rolling census was implemented [31]. The results in 2013 revealed the method's failure and inadequacy for the Israeli scenario (because of the dependency between the sampling frame and the registration in the population registry). This disastrous experience paved the way for the next census, which will be held in 2022, and reinforced the need to take a number of steps that would move the ICBS closer to an administrative census.

The census of 2022 is designed to sample 7% of the population whose usual residence, according to administrative data, is in Israel. Those sampled choose one of three data collection modes: web, telephone, or face-to-face interview to fill out a questionnaire for themselves and for all members of their household. The population estimates will be calculated by comparing the registered address to the actual address given on the census questionnaire as the usual resident address [28].

Before we go into detail about the activities that will help the ICBS make the transition to an administrative census, we clarify the constraints in the population register that are preventing us from doing so.

5.2 The Population Registry's limits

Israel's demographic registration has existed since the country's founding in 1948. Citizens and those who have been granted a long-term residence permit are given a unique ID number and are included into the registry. The usage of the ID number for receiving services from the public and private sectors is fairly common.

Along with ID, Names, Addresses, Date-of-Birth, Sex, Family Ties, Marital Status, Religion, Nationality, Immigration Date, Country of Birth, Country of Father's Birth, and Border Control Information (departures and arrivals in Israel) are all features of the Population Registry that allow ICBS to conduct demographic statistics.

The population register has a variety of restrictions that make transitioning to an administrative census challenging [32]. 1) Failure to capture emigration: The population register includes people who emigrated from Israel and still live abroad or have subsequently died abroad without being de-registered; 2) Failure to capture internal migration: In past censuses, it was reported that approximately 20% of the population was registered in one statistical area but resided in another; 3) People without an official address: A large portion of the population, particularly in Arab towns, is registered in the town but does not have an address that can be geocoded; and 4) Incomplete coverage of usual residents: Foreigners whose usual residence is in Israel but are not recorded in the population register due to its inclusion rules.

5.3 The transition to population estimates based on registry data after the 2022 Census:

7% of those enrolled in the population registration and whose usual residence is in Israel will be sampled in the 2022 census. In each of the statistical areas, sampling occurs in two stages: in the first, the entire family[1] is sampled, and in the second, one of the family members is sampled. Sampled individuals are asked to complete a questionnaire for themselves and all household members using one of three modes: the internet (CAWI), the telephone (CATI), or a face-to-face interview (CAPI). The national population counts are based on the sampling frame plus the addition of foreigners (see below). The findings of comparing the registered addresses of the sampled people to their reported addresses in the questionnaire are used to calculate local population counts. This procedure ensures that inaccuracies in the registered address are accounted for and suitable corrections estimated.

ICBS will receive population estimates for the census day in 2022. It is crucial to note, however, that a number of actions have been developed in parallel with census planning for the benefit of a future administrative census, and the 2022 census is a primary source for analysing these actions. The primary activities intended to promote an administrative census after 2022, as well as the results of the 2022 census and other sources, are listed below.

Emigration Stock: Israel's borders are closed, and everyone who enters or departs the country is tracked by the border control system. For many years, the ICBS has relied on the border control system to

[1] According to the registration, a family is defined as members of the same family who are registered at the same address.

compile statistics on emigrant flows. Ahead of the 2008 census, the ICBS developed an identifiable list of people whose usual residence is outside of Israel (emigrant stock). Following the census, a review of the list revealed that it was of acceptable quality for national population estimates. 7.6% of the list was found to belong to the Israeli populations, and 3.2 percent of the list was missing. As a result, the national population was overestimated by 0.31 percent.

ICBS has updated the procedure for the 2022 census by controlling the list based on emigrant flows. A new departure is a person whose usual residence was in Israel at the determining departure (the determining departure is the departure in which they have accumulated continuously 90 days abroad) and who has spent at least 275 days abroad in the year since the determining departure day. A returning person is a person who was found in the emigrant stock before the determining return (the determining return is a return in which he has spent 90 continuously days in Israel) and has spent at least 275 days in Israel in the year since the determining return day .

In 2020, the corona crisis provided an opportunity to evaluate the emigrant stock. We computed the percentage of people who took a Covid19 PCR test in 2020 from two groups: those who have their usual resident in Israel and those who are emigrants. Only 3.5 percent of individuals living abroad took the test, compared to 40 percent of those with a usual residence in Israel. Because all Israelis living abroad who visited or returned to live in Israel during the crisis were compelled to take the test when they arrived in Israel, the percent of emigrants who took the test is still high, but crucially considerably lower than for usual residents.

Differences between registered and real addresses: Finding additional sources of data that include addresses other than those listed in the registry is one of the key challenges of local level estimates. Because of the presence of these sources, a model for selecting one of these addresses can be developed. We have examined four sources so far: municipal tax payments, electricity payments, student registration and employment registration.

We developed a model for selecting one of the addresses based on a census pilot conducted in 2017. Machine learning tools are used to create the model, which is based on real data from the pilot. The following are the most important findings: 1) About 30% of the population has at least one additional address not shown in the population register; 2) for the 97% of the population, one address (including those with only one address) is the usual resident address; 3) If we choose the address with the highest probability, the model predicts the correct address 92.5 percent of the time, and 90.7 percent of the time if we choose the address with a probability >= 0.5; 4) When we applied the model to the Labor Force Survey and the Social Survey, we got similar results.

The results have significantly improved the situation, but the model still needs to be refined to get closer to the possibility of a correct address selection (97 percent). As a result, fresh sources of addresses must be added, while the model must also be improved and tested in a huge database from the 2022 census.

People who reside in institutions for a long or short amount of time are one source of the disparity between the registered address and the usual address. Typically, these individuals do not alter their registered address to the institution's address. These individuals' addresses will not be discovered in the above-mentioned address sources [32]. As a result, the address selection model will not be able to fix the issue. As a result, another solution must be developed and implemented.

The ICBS performed a census of institutional inhabitants every two years after the 2008 census in order to be more precise in annual population estimates. This solution does not meet our needs in terms of transitioning to an administrative census. As a result, it is determined that after the 2022 census, the CBS will conduct an annual census of institutional residents.

People without an official address: The address selection model provided above does not apply to people whose address in the population registration only comprises the locality of residence, not the street and building. This population prevents us from categorizing people into statistical areas. We allow residents to answer names of unofficial neighbourhoods that we have been able to correlate with statistical areas in the censuses of 2008 and 2022. Because only administrative data should be utilized in administrative censuses, this solution is not suitable.

The problem is most prevalent in Arab towns where there are no official addresses or when residents prefer not to use official addresses. The approach is to encourage residents to utilize official addresses or to locate administrative sources that can anchor the address. We look at two major sources: 1) the use of a municipal tax payment file, which often includes an anchorage of the block and parcel of land where the apartment is located; 2) the use of electricity payments, which provide the exact, coordinates of the unit.

Foreigners: Data on foreigners in Israel is kept by the Population and Immigration Authority which include basic demographic data (age, sex, country of citizenship, marital status and address) and visa data. These figures are based on the border control system as well as other data systems that govern foreigners in Israel, such as a foreign worker system. The biggest issue for foreigners in Israel is that data is available at the national level, but not at the local level [32].

The ICBS excluded foreigners living in Israel in the 2008 census, and population estimates since then has made no mention of them. The ICBS planned to conduct an administrative census for foreigners in 2022 that would address demographic variables without requiring respondents to complete a socioeconomic questionnaire. This census is possible because of collaboration between the ICBS and the Population and Immigration Authority, which has started providing foreign data to the ICBS along with address data, and the authority occasionally supplements the address data in support of the census.

Socio-Economic Data: This paper has concentrated on the shift to administrative data for population counts rather than the gathering of administrative data on social-economic concerns. It is crucial to highlight, however, that the ICBS is also planning to collect administrative data on socio-economic issues. To this purpose, the ICBS is compiling administrative data for the socioeconomic sector, including the Registry of Dwellings and Buildings, the Business Registry, the Employment Registry, the Education Registry, and others.

5.4 Summary

It is clear that the ICBS has made significant progress in the transition to an administrative census, particularly in the area of population counts. The use of a broad database that will be obtained in the 2022 census is expected to allow us to evaluate the actions mentioned in section 5.3. We believe that evaluating these actions and adapting the models to different populations will allow us to move to an administrative census within two to three years. The procedure will be completed later with the addition of the socioeconomic component.

According to ICBS experience, transitioning to an administrative census necessitates a number of considerations that are pertinent to all countries interested in doing so:

1. The transition from a traditional census to an administrative census is a lengthy process that necessitates an examination of the administrative sources accessible to the national statistics office.
2. The moment at which an administrative census can be conducted should be examined. This does not imply that the transition is completed at this moment. The inference is that we have arrived at a point where the users of the data are satisfied with the quality of the data obtained. The data must be improved after the shift to an administrative census by incorporating new data sources and re-examining the procedures.
3. It should be noted that on the side of the expected profits from an administrative census (such as: reduction of costs, reducing the burden, frequency and availability of census data) sometimes in the transition to an administrative census we will have to pay a price. A price that is reflected in changes to the timeliness of the data and a price in reducing the traditional issues

that are associated with a traditional census. Some of the data we lose can be mitigated by conducting dedicated surveys or finding a replacement for it in administrative data.

These are considerations that all countries have in common, so continuing cooperation and sharing of experience is an important part of promoting administrative censuses in each country.

6. Final remarks

It is possible to identify from the above two major alternative frameworks for census-like population statistics, which are currently being developed internationally by countries that lack the population register structure existing in Nordic countries, where the focus is population counts with basic demographics and detailed location.

One approach focuses on the replacement of the basic census enumeration by administrative registers, such that sufficiently detailed population statistics can be produced in combination with coverage survey(s), including using census temporarily as the largest possible coverage survey. Internationally, this approach was pioneered in Israel in 2008. The developments in Italy follow a similar structure, where the PBR is constructed from imperfect registers and estimation of the basic demographic structure is supported by surveys and modelling work. In New Zealand, an integrated population dataset is built around a variety of administrative sources in the absence of population register, where currently the population census figures as the largest possible coverage survey.

Taking this approach, over time, due to improvements of the administrative registers and the survey and estimation methodology, the coverage surveys can be expected to have reduced sample sizes and perhaps reduced regularity. However, to be able to support sufficiently detailed statistics, there will be a limit to how small the sample sizes can be reduced to. It seems reasonable that such a set of *Register Surveys* would require the largest sample across all the programmes of an NSO.

An alternative approach, *fractional counting* [33] provides a unified framework to a fully model-based approach to population statistics. It moves away from the concept of a 'census' being a list of the members of the population with their unique geography and attributes, to the concept of estimating outputs from the available data. Starting from an extended population dataset (EPD) with negligible under-coverage errors, each EPD person is assigned, successively, a probability of belonging to the target population, a probability of living at an unknown address, and a vector of probabilities of living at one of the known addresses. (Over time, it may be necessary to introduce further refinement of this template, as some of the discussion of the work at ICBS seems to suggest.) In the two existing examples, the fractional counters are either set heuristically [34] or based on a predictive model fitting to the last census data that is held fixed subsequently [35]. In addition, three modelling and updating approaches can be identified in the literature.

- Prediction modelling updated by coverage surveys that provide labelled sample observations of in- and out-of-scope persons [36]. The coverage survey may be yearly or less frequently, and the sample size may be relatively large or small.
- Latent class modelling, where the in- and out-of-scope status of a person is treated as a latent variable that is unobserved [37,38]. A notable difference to prediction modelling is that labelled observations are not required, and the model can be fitted to population registers directly – hence, updated over time.
- Trimmed log-linear modelling, where the chosen population registers are trimmed to reduce the erroneous enumerations therein to a negligible extent and a log-linear model [39] for capture-recapture data are then fitted to the trimmed registers. See [40,41] for the basic setting with two lists. The approach is applicable whether or not labelled observations are available by surveys or censuses. It is also possible to extend the approach to accommodate erroneous enumerations, if at least one of the lists is free of such errors [42,43].

To establish the model-based fractional counting system as an alternative to the Register Survey approach above, three key methodological solutions are needed:

- completely register-based population size estimates using statistical models,
- audit sampling inference [44,45] of the register-based population size estimates,
- updating of the underlying fractional counters that are coherent with the population size estimates at more aggregated levels, which can make use of the audit sample as well as other relevant register and on-going survey data.

In particular, provided the system of fractional counters has the adequate quality for producing population statistics on the continuous basis, audit sampling inference can be used to estimate the errors of the model-based statistics, i.e. instead of producing the population size estimates as in the Register Survey approach. This allows one to further reduce the sample sizes and frequency. A simple calculation below illustrates why auditing can reduce cost.

Let θ be the proportion of population units among the EPD persons, and μ that corresponding to the register-based statistics. Suppose audit sampling aims at an unbiased estimator of $(\mu\text{-}\theta)^2$, denoted by mse(μ), and let $\hat{\theta}$ be an unbiased estimator of θ based on the same sample. Based on a technical result in [45], under simple random sampling from the EPD, we have

$$SE[mse(\mu)] \approx 2\,|\mu - \theta| \cdot SE(\hat{\theta}) < \; SE(\hat{\theta})$$

where the inequality holds because one may assume$|\mu\text{-}\theta|$<0.5 without losing practical relevance, i.e., μ is closer to 1 than 0 if $\theta > 0.5$ or closer to 0 than 1 if $\theta < 0.5$. Therefore, the standard error of the target estimator for auditing is always smaller than that of Register Survey aimed at $\hat{\theta}$, such that auditing inference requires a smaller sample size.

It is always tempting but precarious to predict the future. In principle, short of being able to count a Central Population Register directly (as the Nordic countries have done for over 3 decades), the register-based fractional counting (by statistical models, instead of simplistic rule-based categorical classification) in combination with audit sampling inference creates an integrated framework for making an efficient and sustainable system for population estimation. Above all, it urges a conceptual change to census that is visualised as a list of unique population units.

References

[1] Statistics NZ. Introduction to the Census; 2001 [cited 2022 Feb 24]. Available from Stats NZ Store House - Stats NZ Store House (oclc.org).

[2] Statistics NZ. Transforming the New Zealand Census of Population and Dwellings: Issues, options and strategies; 2012 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[3] Statistics NZ. Integrated Data Infrastructure; [cited 2022 Feb 24]. Available from www.stats.govt.nz

[4] Statistics New Zealand. Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project; 2014 [cited 2022 Jun 13]. Available from www.stats.govt.nz.

[5] Bycroft, C, Miller, S, Gath, M, Matheson-Dunning, N, Simpson, K, & Das, S (2021). The quality of administrative data for census variables: Strengths, limitations, and opportunities; 2021 [cited 2022 Jun 13]. Retrieved from www.stats.govt.nz.

[6] Statistics NZ. Experimental population estimates from linked admin data: 2017 release; 2017 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[7] Statistics NZ. Overview of statistical methods for adding administrative records to the 2018 Census dataset; 2019 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[8] Statistics NZ. Data sources, editing, and imputation in the 2018 Census; 2019 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[9] Statistics NZ. 2023 Census: High Level Design; 2021 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[10] Statistics NZ. Experimental *administrative population census*; 2021 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[11] Statistics NZ. Experimental population estimates from linked administrative data; 2018 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[12] Statistics NZ. Experimental administrative population census: Data sources and methods; 2021 [cited 2022 Feb 24]. Available from www.stats.govt.nz

[13] Abbott O, Tinsley B, Milner S, Taylor AC, Archer R. Population statistics without a Census or register. Statistical Journal of the IAOS. 2020; 36: 97-105

[14] Office for National Statistics. Statistical design for Census 2021, England and Wales [cited 2022 Mar 01]. Available from: https://www.ons.gov.uk/census/censustransformationprogramme/censusdesign/statisticaldesignforcensus2021englandandwales

[15] Office for National Statistics. Population estimates for the UK, mid-2020: methods guide [cited 2022 Mar 01]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/populationestimatesfortheukmid2020methodsguide

[16] Office for National Statistics. Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016 [cited 2022 Mar 01]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21

[17] Zhang, LC. On provision of UK neighbourhood population statistics beyond 2021. arXiv [Preprint]. 2020 [cited 2022 Feb 28]. Available from: https://arxiv.org/abs/2111.03100

[18] Office for National Statistics. Measuring and adjusting for coverage patterns in the admin-based population estimates, England and Wales: 2011 [cited 2022 Mar 01]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/measuringandadjustingforcoveragepatternsintheadminbasedpopulationestimatesenglandandwales/2011

[19] Office for National Statistics. Admin-based ethnicity statistics for England, feasibility research: 2016 [cited 2022 Mar 01]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/adminbasedethnicitystatisticsforenglandfeasibilityresearch/2016

[20] Yang, Shu, and Jae Kwang Kim. 2016. "Fractional Imputation in Survey Sampling: A Comparative Review." *Statistical Science* 415-432

[21] Office for National Statistics. Indicative uncertainty intervals for the admin-based population estimates: July 2020 [cited 2022 Mar 01]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/indicativeuncertaintyintervalsfortheadminbasedpopulationestimatesjuly2020

[22] Wolter KM. Some coverage error models for census data. Journal of the American Statistical Association.1986; 81:338-346.

[23] Pfeffermann D. Methodological issues and challenges in the production of official statistics. Journal of Survey Statistics and Methodology. 2015; 3:425–483.

[24] Zhang LC. A note on dual system population size estimator. Journal of Official Statistics. 2019; 35 (1):279-283.

[25] Zhang LC, Dunne J. Trimmed dual system estimation. In: Bohning D, Van Der Heijden PGM, Bunge J, editors. Capture Recapture Methods for the Social and Medical Sciences. CRC Press; 2017. p. 239-259.

[26] Chieppa A, Gallo G, Tomeo V, et al. Knowledge discovery for inferring the usually resident population from administrative registers. Mathematical Population Studies. International Journal of Mathematical Demography. 2018; available from: http://www.tandfonline.com/loi/gmps20

[27] Bernardini A, Cibella N, Gallo G, et al. Empirical evidence for population counting: the combined use of administrative sources and survey data. ESS Workshop on the use of administrative data and social statistics. Valencia. 2019; available from https://ec.europa.eu/eurostat/cros/system/files/gerardo-gallo_empirical-evidence-population-counting_istat.pdf

[28] Israeli Central Bureau of Statistics. From physical area to virtual lists: Toward an administrative census in Israel. Group of Expert on Population and Housing Censuses, Economic Commission for European - Conference of European Statisticians, Twentieth Meeting: Geneva, 26-28 September 2018. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2018/Meeting-Geneva-Sept/ECE.CES.GE.41.2018.21_Item_2_YT.pdf

[29] Israeli Central Bureau of Statistics. Small Area Estimation to correct for measurement errors in big population registers with application to Israel's census. Group of Expert on Population and Housing Censuses, Economic Commission for European - Conference of European Statisticians, Twenty-first Meeting: Geneva, 18-20 September 2019. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2019/mtg1/G1921248.pdf

[30] Israeli Central Bureau of Statistics. Quality Assessments of the 2008 integrated census. Group of Expert on Population and Housing Censuses, Economic Commission for European - Conference of European Statisticians, Twelfth Meeting: Geneva, 28-30 October 2009.

https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2009/12.e.pdf

[31] Israeli Central Bureau of Statistics. The First Round of the Rolling Integrated Census in Israel – Methodology, Results and Flaws. Group of Expert on Population and Housing Censuses, Economic Commission for European - Conference of European Statisticians, Seventeenth Meeting: Geneva, 30 September to 2 October 2015.

https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2015/mtg1/CES_GE.41_2015_4-Israel_rev.pdf

[32] Israeli Central Bureau of Statistics. Estimation of the Total Population in the 2020 Integrated Census in Israel. Group of Expert on Population and Housing Censuses, Economic Commission for European - Conference of European Statisticians, Nuneteenth Meeting: Geneva, 4-6 October 2017. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/WP27_ENG.pdf

[33] Zhang, L.-C. Complementarities of Survey and Population Registers. Wiley StatsRef: Statistics Reference Online; 2022, to appear.

[34] Zhang, L.-C. On provision of UK neighbourhood population statistics beyond 2021. arXiv; 2020. https://arxiv.org/abs/2111.03100

[35] Tiit, E.-M. and Maasing, E. Residency index and its applications in censuses and population statistics. Eesti statistika kvartalikri. (Quarterly Bulletin of Statistics Estonia). 2016;3/16:41-60. http://www.stat.ee/publication-2016_quarterly-bulletin-of-statistics-estonia-3-16.

[36] LCSB. Method Used to Produce Population Statistics. Central Statistical Bureau of Latvia; 2019. https://www.csb.gov.lv/sites/default/files/data/15_04_2019_Iedz_Metodologija_ENG.pdf

[37] Di Cecco, D., Di Zio, M., Filipponi, D. and Rocchetti, I. Population size estimation using multiple incomplete lists with overcoverage. Journal of Official Statistics, 2018;34:557-572.

[38] Baffour, B., Brown, J.J., and Smith, P.W.F. Latent Class Analysis for Estimating an Unknown Population Size – with Application to Censuses. Journal of Official Statistics, 2021;37:673–697.

[39] Fienberg, S.E. The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika, 1972;59:409-439.

[40] Zhang, L., & Dunne, J. Trimmed dual system estimation. In D. Bohning, P. G. M. van der Heijden, & J. Bunge (Eds.), Capture Recapture Methods for the Social and Medical Sciences (pp. 239-259). (Chapman & Hall/CRC Interdisciplinary Statistics). CRC Press; 2017.

[41] Dunne, J. The Irish Statistical System and the emerging Census opportunity. Statistical Journal of the IAOS, 2015;31:391-400.

[42] Zhang, L.-C. Log-linear models of erroneous list data. In Analysis of Integrated Data, eds. L.-C. Zhang and R.L. Chambers. Chapter 9, pp. 197-218. Chapman & Hall/CRC; 2019.

[43] Zhang, L.-C. On modelling register coverage errors. Journal of Official Statistics, 2015;31:381-396.

[44] Zhang, L.-C. Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. Journal of the Royal Statistical Society, Series A, 2021;184:571-588.

[45] Bernardini A., Cibella, N. and Solari, F. A Statistical Framework for Register Based Population Size Estimation. 2022: *Internal report, Istat.*

Table 1. Over and under coverage of PBR from administrative data, based on Experimental DB with coherent signs/PBR, period 2017-2018.

Outcome Description		**Population Counts**
Correctly reported in BPR		**59.197.785**
National under coverage		**290.049**
National over coverage		**1.105.164**
Grey area	Only in BPR	197.621
	Only in AIDA	45.295
		242.916
Not well collocated in BPR		**20.423**
Not eligible for population counts	In AIDA with weak signs of life	**1.410.497**