Rolling Integrated Census in Israel

Dan Ben Hur* and Luisa Burck*
Central Bureau of Statistics, Jerusalem, Israel louiza@cbs.gov.il

Abstract

The 2008 Census of Population was the first Integrated Census (IC) conducted in Israel. The IC was designed to provide population counts and estimates of population characteristics by combining information from the Population Registry (PR) with data obtained from two interlaced sample surveys in the field. Israel's PR cannot substitute for a census as it is not coterminous with the list of persons defined as comprising the de jure census population. The IC coverage estimates at statistical area level are based on the methodology of Dual System Estimation (capture-recapture models), and represent an extension of the classical model for estimating undercoverage in census data; the extension accommodates overcoverage or "false captures" in the administrative data. The undercoverage sample (the U-sample) of approximately 17 percent of the population, serves to estimate the undercoverage parameter as well as obtain detailed demographic, social and economic information. The overcoverage sample (the O-sample), a list of people drawn from the PR, enables to evaluate the validity of the addresses and thus estimate the overcoverage parameter. Based on these two parameters' estimators, we assigned a census weight for each record (person) in PR; a weight which reflects the number of persons it represents in the population. Thus, the population estimate for any population group is the sum of the census weights assigned to its members. The conception behind the Rolling Integrated Census is to establish a multi-year work program that will provide ongoing detailed geographical population counts and estimates of population characteristics. Although the basic methodology of the Rolling Census is based on 2008 Census methodology, some modifications and enhancements concerning to sample design, locality level estimates and the coherency of the estimates have been introduced.

Key Words: Register-based census, homogeneous estimation groups, dual system estimation.

1. Introduction

Israel has conducted six population censuses, in 1948, 1961, 1972, 1983, 1995 and 2008. All Israeli censuses except the last were conventional censuses. In a conventional census, the goal is to physically enumerate every person, either directly or by proxy through member of their household. The questionnaire for the 100 percent of the enumeration contains a small number of basic demographic questions and a longer questionnaire is administered to a 20 percent sample of the households. The 2008 Census of Population was the first Integrated Census (IC), based on administrative sources augmented by survey data for estimating coverage errors. The "best" official list for census purposes is the Israeli Population Registry (PR) and each person in the Population Registry has a unique ID number. In theory, Israel's PR could provide the same information as the short census form: age, sex, address, place

of birth, date of immigration, religion, marital status, kinship relations, etc. The PR cannot substitute for a census: first, the PR is not coterminous with the list of persons defined as comprising the de jure census population. It contains persons who are not part of the de jure population, in particular emigrants who no longer live in Israel and it does not include persons lacking ID numbers who are residents in Israel continuously for a year or longer, legally or illegally. Second, the geographical information in the PR, and in particular the addresses, is of poor quality: approximately one fifth of the persons in the PR are listed at addresses other than where they actually live. Moreover, the PR doesn't include the socio-economic information obtained on the census long form.

The conception behind Israel's IC is simple: instead of a complete field enumeration use the Improved Administrative File (IAF), i.e. PR augmented by other administrative files, as the basis for population estimates and correct these counts using the information on addresses from two interlaced surveys. The first, based on area sample, serves to estimate the undercoverage parameter (the U-sample), as well as obtain detailed demographic, social and economic information. The second, a list of people drawn from the IAF, focuses on the overcoverage parameter (the O-sample).

2. The Coverage Model

The main goal of IC is to provide reliable population counts for small geographic and demographic subgroups. To achieve this, a census list (IAF) of all people in the population at their residential areas, is created. The IAF coverage errors are defined with respect to these areas. For a given area, undercoverage is due to people who live in the area but are listed elsewhere. Overcoverage refers to people who are listed in the area but live elsewhere.

The well-known dual-system multinomial model (Wolter, 1986) provides an estimate of undercoverage in the census list. It is based on two independent enumerations of the population. The first enumeration is the census count and the second is typically based on a sample of geographical areas, called here enumeration areas (EAs). A major difference between the IC and a traditional census is that the traditional approach assumes that there is no substantial overcoverage in the two enumerations. This might generally be true for an area-based enumeration with proper edits and checks of the data. However, the IAF might include, on average, as much as 25% extraneous records for a given area. To handle this problem, Glickman et al. (2003a, 2003b) extended the dual-system model to accommodate overcoverage in the first enumeration.

Let N be the number of eligible people in an area E. The objective is to estimate N. For a person $k \in E,$ let $\mathbf{Z}(k) = (Z_{11}(k), Z_{12}(k), Z_{21}(k), Z_{22}(k))$ be a multinomial random variable indicating the enumeration status of an eligible person k in two independent lists of E: enumerated twice, only in the first list, only in the second list and not enumerated in either list. It is assumed that the random variables $\mathbf{Z}(k)$ are identically and independently distributed with $\mathbf{Z}(k) \sim \text{Mult}(p_{11,} p_{12,} p_{21,} p_{22})$. Define $p_{1+} = p_{11} + p_{12}$ and $p_{+1} = p_{11} + p_{21}$ to be the probability that a person is enumerated in the first and second lists, respectively. Note that we assume here that all persons in the first list have the same enumeration probabilities and similarly for the second list (the homogeneity assumption). In addition, define $Z_{1+}(k) = Z_{11}(k) + Z_{12}(\text{k})$ and $Z_{+1}(k) = Z_{11}(k) + Z_{21}(k).$ Let $X(k)$ be equal to one if a person k listed in the IAF at E lives outside E. Assume

that $X = \sum_{k \in E} X(k)$ is independent of $\mathbf{Z}(k)$ and has a Poisson distribution with parameter λN. Suppose that the area *E* is partitioned into M EAs. A simple random sample of m enumeration areas is selected. The U-sample *S* comprises all eligible people who live in the sampled EAs, and the O-sample *S′* those listed in the IAF in the same areas. Under the above independence and homogeneity assumptions unbiased estimators of *N* and of the coverage parameters are given by

$$\hat{N} = \frac{Z}{\hat{p}_{1+} + \hat{\lambda}}, \quad \hat{p}_{1+} = \frac{\sum_{k \in S} Z_{11}(k)}{\sum_{k \in S} Z_{+1}(k)}, \quad \hat{\lambda} = \frac{\sum_{k \in S'} X(k)}{\sum_{k \in S} Z_{1+}(k) / \hat{p}_{1+}} \quad (1)$$

where $Z = \sum_{k \in E} Z_{1+}(k) + X(k)$ is the observed total number of people listed in area E in the IAF. The linear approximation for the asymptotic variance of $\hat{N}$ is given by

$$\mathrm{Var}(\hat{N}) \approx N\,[o_{1+} o_{+1} + \tfrac{M-m}{m}\{(1-r)r - o_{1+}(1-r-p_{+1}^{-1})\}] \quad (2)$$

where $o_{\square} = (1-p_{\square})/p_{\square}$ is the odds of not being enumerated in count $\square$ and $r = p_{1+}/(p_{1+} + \lambda)$ is the "shrinkage" parameter. The variance reflects model errors as well as sampling errors.

Based on 1995 census data and three field tests carried out pre-IC, it was decided to divide the population of each SA into four homogenous groups with respect to the coverage parameters (based on CART). The coverage model was applied for each estimation group and separate estimators of the coverage parameters at the SA level and locality level were computed. At the end of the estimation process a census weight was assigned to each record in IAF. Thus, the estimated census population for any given geographic and demographic subgroup could be obtained as the sum of the weights of people belonging to the subgroup in IAF.

3. The Coverage Samples in 2008

Since the 1961 census, the CBS has used a system for hierarchical division of urban localities with more than 10,000 residents into geographical-statistical areas (SAs). The SAs system covers the total population of Israel, and is updated before each census. On average, SAs comprise 3000-4000 people. We will hereinafter use the term SA to stand for a statistical area, where it is defined, or a locality, otherwise. The aim of the IC is to provide population estimates by age and sex subgroups within statistical areas.

In preparation for an IC, the country was divided into 53,220 enumeration areas (EAs); each EA included on average 170 people (about 50 households). The EAs are nested within SAs and are used as the sampling unit for the two interlaced coverage surveys. All records in IAF were also geocoded and then clustered by EAs.

The sample for the first IC was planned to comprise about one-fifth of the population, similar to the "long form" sample in the 1995 census. In the IC, the sampling fraction within SAs varied, as opposed to the uniform systematic sampling of every fifth household used in 1995. A random sample of EAs was selected for each SA (total of 9,420 EAs) and all households within the sampled EAs were enumerated. The O-sample consisted of the individuals from IAF, registered in the same EAs sampled for the U-sample.

The 2008 IC made extensive use of the record linkage system (exact and probability matching) especially developed for IC, (Yitzkov and Azaria, 2003). The record linkage process was applied at two different time points: first, people enumerated in the institutions were linked to IAF. Second, the respondents of the U-survey were

matched to IAF and only the remainder of the people in the O-sample were traced and interviewed by phone to determine their status.

4. The Integrated Rolling Census in Israel

Following IC 2008, CBS considered a range of options for the following censuses. First, a full register-based census was ruled out as the results of IC 2008 clearly indicated that the address information in IAF would lead to large biases in the estimators of some socio-demographic subgroups. An organizational decision to adopt a rolling census design, similar to the one implemented in France, was made. In general, two main advantages to a rolling census: spreading the costs much more evenly (maybe also in a flexible way) over a decade and maintaining staff with expertise in census methodology, technology and logistics. For Israel, the rolling census also meant taking constantly advantage of the ongoing household surveys whose number climbed up as Israel became a member state in OECD. This is of significant importance as in the statistical system of a small country, as Israel, relatively large samples in proportion to the population have to be used to attain significant results. In addition, CBS decided to invest in activities supporting census operations that will lead eventually to full register-based census with enriched data from surveys and administrative files.

Primarily, efforts were put into constructing a building and dwelling register that Israel lacked. The building and dwelling register is intended to serve as the sampling frame for all household surveys, thus allowing synchronization between different surveys and wider geographical coverage while reducing the response burden. To achieve these goals, the municipal tax files concerning municipalities and local councils were enriched, revised, standardized and geocoded and there is still ongoing work to improve the quality of the information in these files and especially those comprising localities within regional councils. Furthermore, almost all census questions are harmonized with the ongoing household survey questionnaires and some questions concerning the "reference day" are incorporated in the ongoing surveys.

5. Principles of the Integrated Rolling Census in Israel

5.1. Preserving the methodology of 2008 IC

The conception of the Integrated Rolling Census (IRC) has remained similar to IC 2008: instead of a complete field enumeration use the IAF as the basis for population estimates and correct these counts using the information on addresses from all household surveys (including the two census coverage surveys). The dual system methodology is maintained for the Integrated Rolling Census (IRC) although some special issues may be addressed differently (for instance, a full enumeration of Bedouin population).

5.2. Census of Institutions

A full enumeration of institutionalized population is carried out in a two-year cycle: in the first year the register of institutions is updated and the following year of the cycle, a two stage stratified sampling based on 100% demographic data is performed. First, in each stratum, a sample of institutions is drawn and at the second stage a systematic random sample of the people residing in the selected institutions is drawn. The aim is to collect 100% of the short form every two years and the long form from 50% of the institutions with 10% of the institutionalized population over a decade.

5.3. Complete enumeration over a decade

The sampling design for the census aims to give a complete enumeration of all areas over the cycle of a decade. For large localities with 20 or more SAs, the geo-demographic estimates are computed on an annual basis while the information for smaller localities and SAs within large localities are updated over a ten year cycle. As mentioned above the complete enumeration of the institutionalized population is carried out in a two-year cycle.

5.4. Integration of household surveys

For census purposes, all localities in Israel are allocated into two strata: Localities with 20 or more SAs, and localities with less than 20 SAs. For localities in the first stratum, locality level undercoverage parameters are based on ongoing household surveys as the samples in these localities are large enough to produce reliable estimates. For the estimation of SA level undercoverage parameters, a census-dedicated sample is drawn in addition to the ongoing surveys: the rotation of this sample leads to 100% of SA coverage with 10% of the households interviewed over a decade (10% of the SAs are selected on an annual basis and 10% of the households within each SA are interviewed). Similarly for smaller localities in the second stratum a census sample of 10% of the localities are drawn on an annual basis and 10% of the households within each SA of the selected localities are interviewed yearly. Note that all U-samples are drawn from the building and dwelling register. Independently, samples of individuals registered in PR as residing in the same localities and SAs are drawn on an annual basis for the estimation of the respective overcoverage parameters (O-sample). The O-sampling fraction is about 11% for large localities and 10% for small localities.

5.5. Annual census population estimates

Until the IC 2008, the population estimates were based on censuses as a baseline and were corrected by the "movements" registered in the population counts and demographic changes over time (births, deaths, emigration, immigration, marriages and divorces) that occurred between censuses. The IRC is planned to produce reliable annual population estimates as a whole, as the annual samples serve as an ongoing evaluation survey for the PR. Also on need, the census-dedicated samples can be adopted to meet special needs such as oversampling some population subgroups. Note that the population estimates depend on the coverage parameters that may be stable over time in most areas (other than those with new construction or those that turn into commercial areas).

5.6. Development of socio-economic registers

For the last four years, the CBS has put tremendous efforts in constructing a building and dwelling register that Israel lacked. Although the construction is not completed yet, the individual municipal tax files concerning municipalities and local councils are standardized and geocoded and contain enriched information on the use of the dwellings and the owners. There is still ongoing work to improve the quality of the information in these files and those concerning regional councils.

By the end of 2015, the CBS intends to complete the building of an education register. The education register uses many sources of information as its input, mainly comprehensive files from all education institutes, the Ministry of Education, the Ministry of Absorption and some information on schooling and education from the past censuses and surveys.

In addition, the CBS has an income register; income from work as an employee or a self-employed person. This register is based on income tax files and enriched by other administrative files containing information on businesses, on benefits/allowances or compensations transferred by The National Insurance Institute, etc. All income

information in IC 2008 was extracted from this register (the only exception is pensions).

6. Current Stage

In the beginning of May 2013, the CBS decided to stop all activities involving data collection for IRC except the census of institutions. The first year cycle of IRC, for reference day December 31, 2011, was launched in the beginning of 2012 and data collection was completed by the end of 2012. At the end of March 2013, the preliminary results revealed large biases in the estimators of the undercoverage parameters and further checks confirmed these results. At the moment, a force task, comprising all people involved in the derivation of the estimates, is carrying out a research to explore and explain the results obtained and all activities involving data collection for IRC are frozen.

References

Glickman, H., R. Nirel and D. Ben-Hur (2003a). "False Captures in Capture-Recapture Experiments with Application to Census Adjustment", paper presented at the 54th Biennial Session of the International Statistical Institute, Berlin, Germany.

Glickman, H., R. Nirel and D. Ben-Hur (2003b). "Estimation of Population Size Based on Contaminated Capture-Recapture Data with Application to Census Adjustment", in preparation.

Wolter, K. M. (1986). "Some Coverage Error Models for Census Data". Journal of the American Statistical Association, 81, pp. 338-346.

Yitzkov, T., and H. Azaria (2003), “Record Linkage in an Integrated Census", Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference.