United Nations International Seminar
on Population and Housing Censuses:
Beyond the 2010 Round
27-29 November 2012
Seoul, Republic of Korea

SESSION 4: Emerging methodologies for the census

THE 2008 INTEGRATED CENSUS IN ISRAEL AND FUTURE CENSUSES

Prepared by

Pnina Zadka & Yael Fienstien
Central Bureau of Statistics
Israel

* This document is being reproduced without formal editing.

The 2008 integrated census in Israel and future censuses

Pnina Zadka & Yael Fienstien
Central Bureau of Statistics, Israel

1. Background

The 2008 census in Israel, like censuses in the past, has a primary goal and secondary goals. The primary goal is producing reliable and updated population estimates for all geographic levels up to statistical area (census tract) and for all localities regardless of size in the country.
There are several secondary goals:

- Producing socio-economic-demographic data at all geographic levels
- Serving as sampling frames for future specified surveys
- Possible linking to other personal information sources
- Reach and detailed data source for future research

In all the censuses conducted in Israel in the past(except for the first) the method used two types of questionnaires, a long form and a short form. Both were designed as self administered questionnaires. The short form was delivered to 80% of the dwellings and the long form was delivered to the other 20% of the dwellings. The short questionnaire included basic demographic questions and the full address. The long form also included a set of questions on housing and socio-economic attributes.

The censuses in the past were conducted in the traditional process, a full coverage of all dwellings in the country. Traditional censuses have some drawbacks:

- Cost- high per capita costs -which rise with each census in real cost and increasing population size
- Declining response rates
- Staleness - Fast rate of census data becoming obsolete due to fast changes in the population
- Dated technology - Emerging new technologies which cause the previous census technologies to become irrelevant.

An immediate and low cost option is a register-based census. Register based censuses were conducted since the late 20th century in some European countries.

Israel maintains a few registers that carry the potential to ground a register based census. Nevertheless, these registers suffer from flaws that require processes that will improve the data quality. This led the Israeli Central Bureau of Statistics (ICBS) to develop a method that will make use of the registers combined with specially designed processes which will complete and improve the register.

The most comprehensive register available in Israel is the Central Population Register (CPR). This register contains information on the permanent residents

of Israel since the establishment of the state in May 1948. Each person in the register has a unique Identity number which is granted:

a. At birth in the country or
b. On the first entry to the country with a permanent residency visa or
c. Upon changing residency status during a temporary stay in the country.

The data in the CPR contains all the information required in the past censuses (traditional censuses) in the short form. So it would have been reasonable to anticipate that there won't be any need for the 80% short form process.

This possibility was checked and found to be insufficient due to four reasons:

a. The list of permanent residents in the CPR does not conform with the definition of a usual resident as there are almost 12% of the persons in the CPR who do not reside usually in the country (i.e. they live abroad).
b. The address in the CPR is not the current address of the resident or the address listed is not a full address (for about 20% of the addresses)
c. The information in the CPR relates on to formal family relation and does not depict fully the actual household typology.
d. There in no information on housing or socio-economic characters in the CPR

2. The main principles in the Integrated Census

הכותרת פחות מתאימה

Overcoming the CPR defects/drawbacks/weaknesses may be partially accomplished by using additional available registers. For example, we may delete/ mark as "not belonging to population" people who do not reside usually in the country, by using border control registers and the government allowances register maintained by the National Insurance Institute (the Israeli Social Security Agency)

Other defects which are crucial to the register based census relate to updated addresses.

The CPR contains mainly demographic information and is missing economic, social and housing information ???

The basics of the integrated census are to develop a system which would complete and correct the missing and inaccurate information of the CPR.

As mentioned the CPR, like most registers. is subject to over registration and under registration - residents who are not registered but do live in the country and vice versa, people who are registered but actually live in another country usually.

A variety of administrative sources and two large sample surveys are used to complete and correct the CPR.

3. Correcting the CPR using administrative sources

At the first stage the CPR is corrected using external administrative sources. This process produces the improved CPR (ICPR

a. Border control files and NII files are used to delete persons living abroad
b. Non residents file is used to add persons residing in the country

The main defect in the CPR is, the incorrect addresses. These erroneous addresses amount to almost 20% of the usual residents. About 12% are registered in their right locality but in an incorrect address within the locality. Another 8% reside in a different locality. These estimates are quite stable over the years and were verified again in the pilot census. One of the main groups with incorrect address registration are people residing in institutions (collective housing) such as boarding schools, dormitories and long term care facilities.

The integrated census is designed to correct these inaccuracies in the CPR.

4. The integrated census process

The process is an expansion of the "dual system estimation" for the smallest areas for which the census provides estimates, which are statistical areas (SA) - the equivalent of census tracts.
The **under-count** estimation is based on an area sample. For the under-count sample the country was divided into forty thousand "sampling cells" each containing 50 households on average based on the ICPR and with clear boundaries. The sampling cells were nested within an SA, sampled cells were evenly spread over the SA. This survey was conducted with a computerized questionnaire administered by an interviewer using a laptop. The questionnaire served also to add census information missing in administrative sources. The proportion of people enumerated during the area survey who were not properly registered in the ICPR was used to estimate the under-coverage.
The over-count estimation is based on a sample of "administrative households" drawn from the ICPR registered in the addresses of the sampled cells for the under-count sample wherever possible. An administrative households are a proxy of an actual household. In the pilot censuses taken in 2004 and 2006 it was found that 85% of the administrative households represent the actual household. The survey was conducted as a computerized telephone interview. The proportion of people enumerated during the telephone interview who were not properly registered in the ICPR was used to estimate the over coverage.

Using the two surveys and the record linkage process[1], population was divided into three groups:
a. Found in the address registered in the CPR
b. Found in an address different from the registered address
c. Not found in the registered address

The relations between the three groups are the parameters used for the estimation formulas.

[1] T. Yitzkov and H. Azaria "Record linkage in an integrated census" *Proceedings of the 2003 Conference of the U>S> Federal Committee on Statistical Methodology* (2003)

5. The integrated census statistical methodology

The methodology adheres to the two main goals of the census:

- It updates and improved population estimates
- Socio-economic – demographic characteristics in all geographic resolutions

The first goal, updated population estimates is achieved through the **dual system Estimation model**.
Over-coverage (false captures) - the false inclusion in the classic model of capture/recapture - is represented in the census model by two independent lists: one from the ICPR and the other based on the results from the area sample.

The second goal, socio-economic –demographic characteristics - is achieved through the extended questionnaire used in the under-coverage survey. The responses from the survey are calibrated using weights to the population estimated in each SA[2][3]

N(i) Number of people living in the area to be estimated
U_{11}(i) Number of people living in area i and registered in area i
U_{21}(i) Number of people registered in area i and not living there
U_{12}(i) Number of people living in area i and not registered there

$$P_{1+}(i) = \frac{U_{11}(i)}{N(i)} U_{11}$$

$$\lambda(i) = \frac{U_{21}(i)}{N(i)}$$

$$N(i) = \frac{U_{11}(i) + U_{21}(i)}{P_{1+}(i) + \lambda(i)}$$

The parameters U_{11}, U_{21} and U_{12} are not known and are estimated from the two surveys in the sampled areas.

[2] H. Glickman, R. Nirel and D. Ben Hur, " False Capture in Capture-Recapture Experiments with Application to Census Adjustments" *Bulletin of International Statistical Institute* 54th session, Contributed Papers, Vol. LX, pp413-414

[3] R. Nirel, H. Glickman and D. Ben Hur, "A strategy for System of Coverage Sample for an Integrated Census" *Proceedings of Statistics Canada Symposium* 2003: Challenges in Survey Taking for the Next Decade (2004)

The parameter $\hat{P}_{1+}(i)$ is derived from the Under-coverage survey and the parameter $\hat{\lambda}(i)$ is derived from the Over-coverage survey

In/Out table

ICPR	Under-coverage survey	
	Enumerated	Not enumerated
Registered	$\hat{U}_{11}$	$\hat{U}_{12}$
Not registered	$\hat{U}_{21}$	

N(i) Number of people living in area i (the number to be estimated)

(i)$\hat{U}_{11}$ Number of people that were found in the survey in area i and are registered in area i

(i)$\hat{U}_{21}$ Number of people registered in the ICPR sample in area i and not living there derived from the telephone sample

(i)$\hat{U}_{12}$ Number of people found in area i in the area survey and registered in another address in the ICPR.

$$\hat{P}_{1+}(i) = \frac{\hat{U}_{11}(i)}{\hat{U}_{11}(i) + \hat{U}_{21}(i)}$$

$$\hat{\lambda}(i) = \frac{\hat{U}_{21}(i)}{(\hat{U}_{11}(i) + \hat{U}_{21}(i)) / \hat{P}_{1+}(i)}$$

$\hat{P}_{1+}(i)$ - the estimator of $P_{1+(i)}$

$\hat{\lambda}(i)$ - the estimator of λ (i)

6. The census weight

The over- and under- coverage parameters are calculated for each estimation group (about 12,000 groups for the total country). Each person in the ICPR is assigned a weight based on the estimation group he belongs to. Every person within the same estimation group is assigned the same weight. The weight can theoretically be between 0 and the total number of people in an SA. In practice 95% of the weights run between 0.5 and 2.

The census weight formula is therefore

$$\hat{w}(i) = \frac{1}{P_{1+}(i) + \lambda(i)}$$

7. Census estimates Calibration

The census population estimates, subdivided by geographical breakdown are calibrated to give a total population estimate. These estimates resemble the estimates obtained from the short form in a traditional census
The socio-economic-demographic estimates in each SA are calibrated to the population estimate obtained from the previous process.

Calibration of the weight is performed in an ascending order, first the country total and then for the smaller geographic breakdowns.

8. Disadvantages of the integrated census

a. The integrated census provides only limited information on housing such as condition of the buildings, type of buildings, use of dwellings and housing units, age of building, and average number of dwellings in building.
b. It is not possible to provide data on flexible area, only aggregated data from the SA level and up. No information on buildings.
c. Using the same weights for an estimation group can sometimes distorted differences with the estimation group and

9. The Integrated Rolling Census in Israel -2020 round of censuses

The 2008 integrated census (IC), although reducing the cost substantively, still required extensive organizational, logistic and human resources for the implementation of the two large sample surveys.

The outcomes of the 2008 IC initiated the search for ways to improve the IC methodology and to overcome these deficiencies in the upcoming 2020 round of censuses. An organizational decision was made that the resources allocated for the census will be spread throughout the decade, while the total budget will remain the same. Any savings in the census operation will be invested in census supporting operations which will lead eventually to a full register-based census in the future.

10. Principles and requirements for the upcoming 2020 round of censuses

10.1 Preserving and maintaining the methodology and technology developed for the 2008 census.

- The methodology of dual system estimates that was developed for the evaluation of total error and undercount in censuses (Hogan & Walter, 1988) was extended to include the evaluation of over-count for the integrated census. The 2008 census results proved that the flaws in the CPR were still in a magnitude not tolerable in a full register-based census.

- The technological tools and applications (for data collection, data integration and so on) developed for the 2008 census incorporated many innovations that proved to be functional for the achievement of most of the goals set ahead of time, in particular those regarding timeliness and accuracy. Their costs amounted to more than one third of the total census costs, and they are envisaged as an infrastructure for further development and improvement.

10.2 Preserving experienced manpower and acquired expertise, while minimizing the fluctuations in the organization's manpower and the logistics implications.

- The staff recruited for the 2008 census had almost 10 years of experience in statistical methodology, technology and logistics. In previous censuses most of the staff was lost mainly to other organizations and to other units within the CBS. Most of the experience and professionalism acquired during the planning and execution phases were lost. As specialists in censuses are rare, it was essential to find a way to preserve the leading professionals in the CBS and particularly in the census various units.
- The peaks in resources (human and budget) are troublesome vis-à-vis the flow of the work, logistics within the CBS and budget allocation by the Ministry of Finance. Recruiting and training numerous new staff members is traumatic to any organization. Merely the allocation of experts' time, workspace and work-aids disrupt the routines in the whole organization. Avoiding peaks has its benefits per-se.

10.3 Harmonizing sampling frames used for households and individuals in current surveys.

In the last few years, since Israel got into the accession process to the OECD, there was an expansion in the number of current surveys conducted by the CBS in order to comply with data requirements of the OECD. The size of the population and the increase in population involved in surveys increased the probability of an individual to be sampled. This led to the decision to synchronize Household (HH) samples in order to reduce response burden.

10.4 Developing a geo-coded Building and Dwelling Register (BDR).

During the developing stages of the 2008 IC one of the registers that was lacking was a building and dwelling register. Several options were examined and eventually, the development of a register was abandoned due to budget restraints. Nevertheless the need for a BDR has not vanished, and has even expanded to other statistical domains. Budget planned for future census and staff from the previous census, have been allocated for the development of this register.

10.5 Making use of all data available to increase robustness of census estimates.

Data obtained from administrative sources and from sample surveys are to complete and enrich census data.

10.6 Incorporating and harmonizing census questions in current HH surveys to increase sample size.

Questions shared by censuses and surveys are practically increasing the sample size of the census-survey and therefore improve census estimates.

11. Annual population estimates will be updated by census results.

Annual population estimates were based, in the past, on censuses as a baseline and on the demographic changes registered in the CPR (births, deaths, marriages and divorces - vital statistics events - as well as local and international migration). This process distorted the accuracy of the estimates because of several problems immanent in the data sources:

- Flaws in the census data were kept until the next census corrected them.
- The CPR carries different errors because of late or wrong updates which concentrate in specific areas, like in addresses of interest. For example, the population estimates for new localities that were populated after the census, were based completely on updates of the CPR that are less reliable than the census. These constraints raised criticism on the CBS regarding local population estimates.
- The estimation of institutional population was updated only in the censuses. There was a need to find a procedure which will facilitate annual update for new localities and changes in dwelling regions as well as updating institutional population frequently.
- The upcoming census should address and facilitate these requirements

11.1 Incorporating GPS technology in census operation and in other current HH and business surveys to correct for flaws in addresses.

The expansion of HH surveys conducted by the CBS annually enables the CBS to check and improve geo-coding quality of about 3%-4% of the addresses in the CPR as well as in the BDR, currently under development. The vast use and the price reduction in the cost of GPS devices, led to equipping of the interviewers in the field with GPS devices and obtaining the XY-coordinates for each dwelling unit in the sample. These coordinates will be used as a quality control on the coordinates achieved through address geo-coding and cadastre geo-codes. The equipment will also be used as an aid in navigation in areas without street names and will reduce the amount of non-traceable dwellings.

11.2 Two issues are especially addressed; budget and response burden.

- The decision made with the Ministry Of Finance was to stay within the limits of the original total budget allocated for the 2008 census and that any budget left will be used for developing census supporting data sources and registers. The expectation that budget will be available is based mainly on the evaluation of costs saved by avoiding peaks in the census activities.
- Statistical tests made on the 2008 census data proved that there is no significant increase in the variances of census estimates if a 10% sample is drawn for each survey rather than 17% in the under-coverage survey and 20% in the over-coverage surveys. This led to the

organizational decision to reduce sample size which was crucial to answer the need to reduce response burden inflicted on the population.

11.3 The time and budget savings facilitate the development of three major registers: BDR, Education Register, and Income Register:

- The BDR is based on municipal taxation lists, and when completed, will be used as the sampling frame for all HH surveys carried out by the CBS. Currently the lists vary in quality and some are of insufficient quality and need an extended improvement before they can be incorporated into the register.
- The education register is based on annual files received from all licensed education institutes in the country. The education register will be used to correct and complete education data in the IRC.
- The income register is based on income taxation files and supplemented by data on income received through various social security allowances. It was already used in the 2008 IC to add income information to the census file.

12. Census of Institutions (collective dwellings)

A full census of institutionalized population will be carried out in two years cycles. The first year in each cycle will be devoted to updating the Institutions Register. In the second year of each cycle a two phase enumeration will be carried out; the first will be of a full demographic enumeration of the population in all institutions in the register, and the second will be of a socio-economic sample survey of the institutionalized population. The sample is a two-stage sample; In the first stage 10% of the institutions and in the second 20% of the residents within each sampled institution are sampled. Over a ten-year period, 50% of the institutions will be sampled and 10% of the institutionalized population will be interviewed in the socio-economic survey.

13. The future censuses

The type of census chosen to meet the terms set for the upcoming census was an **Integrated Rolling Census (IRC)**. The main principles of the 2008 census are to serve the IRC. Yet, the operation of the census is to be applied in a ten years period.

14. The sampling methodology

The sampling is based on the scheme developed for the 2008 IC with only few modifications:

- Under-coverage estimation survey- Hierarchical annual three-stage sampling process of municipalities, 10% of the Statistical Areas (SA) within municipalities, and 10% of the dwelling units or buildings in each sampled SA (wherever possible).
- Over-coverage estimation survey – An improved and geo-coded CPR is used as the sampling frame. Sampling 10% of the administrative HHs registered in the sampled SA, amounts to 1% of the total number of HHs each year. Another 1% is sampled each year for the over-count survey from all non sampled SAs and they are used for the over-count estimates on a municipality level.

15. The Labor Force Survey (LFS) as a supporting survey

The LFS, a current monthly survey, conducted by the CBS will be used to support the census coverage and reliability. The expansion of the sample size, the improved geographic coverage and distribution of the survey, as well as the introduction of census questions in the LFS questionnaire, will be used for the estimation of the undercount on the municipality level.

16. Technology

16.1 The Technological tools and applications developed for the 2008 IC are modified slightly to be used for the upcoming years. Bugs discovered in the 2008 IC were also corrected. These minor changes reduced the time and personnel needed for development and enabled launching the first round of the census of institutions in 2010 and the first round in the IRC in 2012.

16.2 The technology and operational procedures developed for the 2008 IC and the IRC were found to be adequate for other HH surveys and were modified to support a full range of ad-hoc surveys which can be implemented with fewer resources (time and staff).

17. Census staff

Most of the experienced personnel that developed the 2008 IC remained in the CBS and was involved in the development of the IRC. The managing staff also remained involved in the census operations.

18. Current stage

18.1 Currently, the CBS is engaged in the data collection phase of the 2011 IRC that was launched in February 2012, and is planned to end no later than November 2012, Reference day is December 31, 2011. The CPR and the BDR were used as the sampling frames for the surveys. The first round of the Institutions census was completed and data was incorporated in the 2011 population estimates.

18.2 Annual population estimates for the SAs will be adjusted based on the updated census results and will be calibrated to the municipality adjusted estimates. The adjustments will be based on the IRC and on the estimated total population, taking into account vital events and migration registration in the improved CPR. The estimates regarding the institutionalized population will be updated every other year for the production estimates of the non-institutionalized population that will be used for the estimation and calibration of the HH surveys.

18.3 The Socio-economic data will be generated for the mid-term of the census period (5 years after the first year in a 10 years cycle). The first full cycle of data collection will be completed in 2022. Nevertheless, preliminary results will be released during the upcoming years.

19. Contemplating Remarks

19.1 The rolling integrated census (IRC) is designed as the beginning of a process aiming toward a full register-based census. It will also serve as an ongoing evaluation survey, to facilitate quality control of the available

administrative sources and registers to support a fully register based census.

19.2 The first steps in the implementation of the IRC seem to meet major goals set for the 2020 census round. It proves to be fully integrated in the organizational operations and mode of management. It reduces the peaks in organizational operations, and saves budget and time used for training inexperienced personnel. The preliminary findings, generated by monitoring the data collection process, show improved quality of data collected, achieved through better and more extensive training of interviewers by experienced staff. The management plans and the data collection technology are implemented in a short period of time, and are running smoothly with no need for a pilot test.

19.3 Yet, since the census is conducted on an ongoing basis, it is dependent on ongoing provision of resources (budget, human resources and administrative data). This dependency can be the Achilles heel of the census in times of crises or instability.

A careful risks management will be required, as done in a new mode of indirect data collection from municipalities, that are currently tested - data submission under cooperation agreement. It addresses the legal and operational issues raised. This mode will also allow for quality control of data upon capture. Hence, municipalities will be involved in improving their own taxation lists in the future.

19.4 Public criticism has already been directed to the IRC: Municipalities and the research community that are used to the data produced in a traditional census would like the next census to provide the very same information.

19.5 These issues should be solved partially by the foreseen advancement in use of statistical registers such as the occupation and industry register (on the individual level), as well as those that are already available: the BDR, education and income registers. The registers may be linked to the improved CPR producing a full census including the core socio-demographic information for the whole population. A core element which will be still missing is the household composition, specifically the informal compositions such as cohabitation and living with a caregiver.

19.6 Another aspect is shared by all censuses and it refers to the need to be attuned to changing technology and to adopt developments needed to better achieve the goals under effective limitations.

19.7 Last but not least is a shortcoming in the concept of the IRC as a census: The socio-economic data is averaged over 10 years rather than being a snapshot. It may cause no problems in static areas, but dynamic areas, in which the population size and their characteristics change often, will have to be dealt with in tailor-made census processes. Moreover, it is expected that globalization processes will further introduce challenges for censuses and even more so for official statistics in general. There is still a need to develop a methodology to pool socio-economic information over ten years period to an average.