Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

A STRATEGY FOR A SYSTEM OF COVERAGE SAMPLES FOR AN INTEGRATED CENSUS

Ronit Nirel, Hagit Glickman, and Dan Ben-Hur[1]

ABSTRACT

The Israeli 2008 census of population is entitled an Integrated Census (IC). It will be based on administrative sources, and augmented by survey data for estimating and adjusting for coverage errors in the administrative data. Two coverage samples will be designed: the U-sample focuses on undercoverage and the O-sample on overcoverage. This paper describes the design of the samples and analyses factors that affect the scope of the direct data collection. The suggested methodology is illustrated for the first IC experiment.

KEYWORDS: Administrative Census; Geocoding; Overcoverage; PES; Record Linkage; Undercoverage.

1. INTRODUCTION

The Israeli 2008 census of population is entitled an Integrated Census (IC). The census will be based on administrative sources augmented by survey data for estimating and adjusting for coverage errors (Blum, 1999). In 1995, the Central Bureau of Statistics (CBS) has conducted the fifth census of population and housing in Israel. It was based on a nationwide area-enumeration (a traditional census) and was followed by a post-enumeration survey (PES). The new approach is expected to improve the quality and timeliness of the estimates, reduce the response burden and, in the long run, increase the census frequency. Cost is expected to reduce and the expenses divided more evenly over time.

The main source used for the IC is the Population Register (PR). Although in general quite reliable, the Register is exposed to coverage errors including overcoverage of emigrants (5-10%) and outdated addresses (10-30%). The problem of undercoverage is well known in census methodology and is traditionally handled by an area-based coverage survey. The additional and new challenge of the IC paradigm is to measure and adjust for sizeable overcoverage. The IC methodology is based on an adjustment of the Register counts through two interlaced surveys. The first, based on an area sample, focuses on undercoverage (the U-sample), and the second, based on a sample of people drawn from the administrative file, focuses on overcoverage (the O-sample). To estimate undercoverage rates, the dual-system model is used, where the first enumeration is provided by the administrative data and the second enumeration by the U-sample. To estimate overcoverage rates, the eligibility and census day address are verified for each person in the O-sample, using fieldwork data as well as administrative information and statistical modeling.

An IC is generated by five sub-processes: creating an integrated administrative file (IAF), collecting sample survey data, linking the two sources, estimating the census weights, and post-evaluating the estimates. Some of these processes are used more than once in a single IC cycle. For example, a linkage between the administrative and survey data is performed after the completion of each survey.

The *administrative process* augments the PR data by other administrative sources. These may include municipal tax files, electricity company files and education system files. The additional files are mainly used to validate or replace addresses in the PR. All addresses in the resulting file are assigned geographical xy-coordinates, an operation called "geocoding" (Blum and Calvo, 2001). During a single IC cycle three versions of the IAF are created:

[1]Central Bureau of Statistics, 66 Kanfey Nesharim St., Jerusalem 95464, Israel, ronit@dev.cbs.gov.il; hagit@dev.cbs.gov.il; danb@dev.cbs.gov.il

- *IAF(s)* provides background information on the population and is used for the design of the samples. It is created about 10 months before Census Day.
- *IAF(l)* is created a short time before the U-sample fieldwork ends (about one month after Census Day) and is used to define the remainder of the O-sample (see Section 3). It is based on a PR updated to several weeks before Census Day.
- *IAF(e)* is the final file, which is based on a PR updated to Census Day. It is used to build the IC estimates and is created about four months after Census Day.

The *sampling process* designs the coverage samples and collects the data. This paper focuses on the design of the U-sample and O-sample and the relationship between the two. Section 3 describes the design of the mainstream part of the samples, and Section 4 analyses the factors that affect the scope of data collection. The mainstream design relies on the ability to geocode the IAF records. Some localities, mainly Arab towns and other small villages, have no addresses in the IAF. Section 5 suggests an alternative approach for these localities.

The *linkage process* matches the IAF and survey data by exact and probability record linkage procedures (Yitzkov and Azaria, 2003). The linkage relies primarily on a 9-digit personal identification number, and is supported by other variables such as name and age.

The *estimation process* starts with estimation of the coverage parameters of an extended dual system model (Glickman et al., 2003a, 2003b). Since this model is the basis for the design of the coverage samples we start by describing it briefly in Section 2. Other elements of this process include diagnostic checks against current demographic estimates and adjustments for small domains of interest. At the end of the estimation process a census weight is computed for each IAF record. Finally, the *evaluation process* assesses the quality of the final estimates using a variety of tools, including an evaluation sample survey.

The first IC experiment was conducted in 2002 in a small town near Jerusalem. Section 6 illustrates the mainstream sample design method using this data. The second experiment is scheduled for November 2004. It will include nine localities in central Israel, comprising about 150,000 residents. Section 7 concludes the paper with a discussion of some additional issues that will be dealt with in the 2004 experiment.

2. THE COVERAGE MODEL

The main goal of a population census is to provide reliable population counts for small geographic and demographic subgroups. To achieve this goal a census list is created, in attempt to count all people in the target population at their residential area. The IC uses an integrated administrative file, the IAF, as its list.

Coverage errors of the IAF are defined with respect to Statistical Areas (SAs, see Section 3). For a given SA, undercoverage is due to people who live in that area but are listed elsewhere. Overcoverage consists of people who emigrated and people who are listed in the area but live elsewhere. The main reasons for differences between an IAF address and a residential address are delayed address updates and benefit considerations (e.g. reduced tax area, school registration area). Note that a discrepancy between an IAF address and a residential address is not considered a coverage error if the two addresses point to the same SA. The IC estimation approach aims to provide direct population estimates for SAs. Estimates at higher geographical levels (e.g., cities) are obtained as the sum of the appropriate SAs estimates. In the following we will concentrate on estimating the population size of a single SA, say *E*. Eligible people living in *E* are called *E*-eligible, and ineligible or eligible people living outside *E* are called *E*-ineligible.

Traditionally, undercoverage of census lists is estimated based on the well-known dual-system multinomial model (Wolter, 1986). The model uses two independent enumerations of the population. The first enumeration is the census list and the second is typically based on a sample of geographical areas, called here enumeration areas (EAs). The IC uses the IAF administrative count and the U-sample area-based count, thus independence is easily obtained. 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 estimate overcoverage, the eligibility of all IAF records listed in the sampled EA's is checked. Estimation is now based on an extended dual-system model that accommodates overcoverage in the first enumeration (Glickman et al., 2003a, 2003b).

Let N be the unknown number of E-eligible people, and suppose that E is partitioned into M EAs of size $N_1,\ldots,N_M$, $\sum_{j=1}^{M} N_j = N$. Let $\mathbf{Z}(j) = (Z_{11}(j), Z_{12}(j), Z_{21}(j), Z_{22}(j))$ describe the number of *E-eligible* people in EA j that are enumerated twice, only in the IAF list, only in the area-based list and not enumerated in either list. Let $X(j)$ be the number of *E-ineligible* people enumerated in the IAF at EA j. It is assumed that $Z(j) \overset{ind}{\sim} Mult\left(N_j; p_{1+}p_{+1}, p_{1+}(1-p_{+1}), (1-p_{1+})p_{+1}, (1-p_{1+})(1-p_{+1})\right)$ independent of $X(j) \overset{ind}{\sim} Poisson(\lambda N_j)$. There are, therefore, two parameters that characterize the coverage of the IAF count: p_{1+}, the rate of E-eligible people (undercoverage parameter), and λ the rate of E-ineligible people (overcoverage parameter). The third parameter, p_{+1}, characterizes the area-based enumeration Homogeneity of the parameters can be approximately achieved by proper demographic stratification.

.

Let S be a simple random sample of m EAs. In addition, define $Z_{1+}(j) = Z_{11}(j) + Z_{12}(j)$, $Z_{+1}(j) = Z_{11}(j) + Z_{21}(j)$. The U-sample and O-sample data provides measures of $Z_{11}(j), Z_{1+}(j), Z_{+1}(j)$ and $X(j)$ for $j \in S$. Information about the non-sampled EA is limited to the total number of people counted in the IAF, $Z_{1+}(j) + X(j)$. Note that $X(j)$ and $Z_{1+}(j)$ are not observed for $j \notin S$. Under the above assumptions, asymptotically unbiased estimators of N and of the coverage parameters are given by

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

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

$$Var\left(\hat{N}\right) \doteq N\left[o_{1+}o_{+1} + \frac{M-m}{m}\left\{(1-r)r - o_{1+}\left(1-r-p_{+1}^{-1}\right)\right\}\right], \tag{2}$$

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

3. DESIGN OF THE COVERAGE SAMPLES

We begin with a description of the statistical areas system. 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 more than 80% of 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 is divided into enumeration areas (EAs). Each EA includes on average 170 people (about 50 households). The EAs are nested in SAs and are used as the sampling unit for the coverage surveys, as described below. The sample for the first IC is planned to comprise about one-fifth of the population (1.3 million people or about 400,000 households), similar to the "long form" sample in the 1995 census. In the IC, the sampling fraction within SAs will vary, as opposed to the uniform systematic sampling of every fifth household used in 1995. A random sample of EAs will be selected for each SA.

The number of EAs in the sample is based on the accuracy requirements and on the level of the coverage parameters. Suppose that we are interested in estimating the size N of a subgroup in a specific SA, where

homogeneity holds. Re-write (2) as $Var(\hat{N}) \doteq N\left(A + \frac{M-m}{m} B\right)$, with $A = o_{1+}o_{+1}$ and $B = (1-r)r - o_{1+}(1-r-p_{+1}^{-1})$.

For a given relative standard error (RSE), $\alpha = \{\mathrm{Var}(\hat{N})\}^{1/2} / N$, the required sample size is

$$m(\alpha) = M \frac{B}{\alpha^2 N - A + B}. \quad (3)$$

To solve (3), estimates of the coverage parameters and of the size of the subgroup are plugged in. However, for given SA and RSE, a different number of EAs may be required for different subgroups. It is customary to choose key estimates for the design or try to protect against "worst case" scenarios. In our case we need to take care of small subgroups with relatively high undercoverage and overcoverage rates, e.g. young people aged 20-30 who tend to move around and do not update their address in the PR. Note that the variance of the estimated size of the whole SA, say $\hat{N}$, which is the sum of the estimates of appropriate subgroups, $\hat{N}_g$, $g = 1, \ldots, G$, is approximately equal to $\mathrm{Var}(\hat{N}) \approx \sum_{g=1}^{G} \mathrm{Var}(\hat{N}_g) \le (N\alpha_{g_0})^2$, where g_0 is the subgroup with highest RSE, which is used to determine m in (3). Therefore, $\mathrm{RSE}(\hat{N}) \le \alpha_{g_0}$.

The sample allocation mechanism is discussed next. For any SA, i, equation (3) determines the relationship between the desired RSE, α_i, and the respective sample size m_i. Let the overall sample size be $\boldsymbol{m} = \sum_i m(\alpha_i)$. For a given $\boldsymbol{m}$, the allocation of the sample to the different areas, which provides a uniform RSE across all areas, is calculated by finding the root α of the function

$$f(\alpha) = \sum_i m_i(\alpha) - \boldsymbol{m} \quad (4)$$

where $m_i(\alpha)$ is derived from (3). If different levels of RSEs are required for different areas, these differences can be expressed as multiples of α, that is $\alpha_i = c_i\alpha$, $i = 1, \ldots, I$. The allocation given by $m_i(c_i\alpha^*)$, where α^* is the root of the function $f(\alpha) = \sum_i m_i(c_i\alpha) - \boldsymbol{m}$, preserves the desired RSE ratios.

To obtain preliminary estimates of the coverage parameters for the design, a detailed analysis of the propensity of the population to be registered correctly in the IAF is carried out. The process involves examination of socio-demographic characteristics such as percentages of children, young people, elderly people, religious people and new immigrants. In addition, rough estimates of the coverage parameters are obtained (a) by matching the 1995 census to the respective PR and (b) from current data on internal immigration.

All IAF records are geocoded and are then clustered by EAs. Records that are not geocoded (incomplete or unknown address) form a separate O-sample stratum, the NG-stratum. There is also a separate U-sample stratum, the NN-stratum, for EAs in new neighborhoods that have practically no people who are registered correctly in the IAF. Generally, the same EAs are selected for both surveys. There are two exceptions: the U-sample may include a sample of EAs from the NN-stratum, and the O-sample may include a sample of records from the NG-stratum. The U-sample comprises all eligible people who live in the sampled EAs on census day. The O-sample includes all people who are listed in the IAF in the sampled EAs. The allocation derived from (4) applies to the common part.

Operationally, the U-sample fieldwork starts one day after Census Day and lasts about 6 weeks. It should be emphasized that under no circumstances will information from the IAF be used during the U-sample fieldwork. This ensures independence between the IAF and U-sample enumerations. The U-sample file (USF) is then matched to the IAF and a list of the remainder of people in the O-sample EAs is created, i.e., the list of people who are included in the O-sample but not in the USF. These people may be either *E*-ineligible, or *E*-eligible people missed by the U-sample enumerators, where *E* is the respective SA. To complete the O-sample fieldwork, all people in the remainder list are traced and interviewed to determine their status. This portion of the O-sample is called the *remainder* of the overcoverage sample (RO-sample). Note that in the RO-sample the enumeration unit is an individual, whereas in the U-sample it is a household. The RO-sample is expected to include on average about 20% of the O-sample. Eligibility and census address on Census Day are verified for both samples

4. FACTORS AFFECTING THE SCOPE OF DATA COLLECTION

In this section we discuss different factors that affect the scope of the direct data collection and are related to the IC process and to the fieldwork methodology. For the U-sample we are mainly concerned with undecoverage of eligible people in the field enumeration. For the O-sample the actual scope of the fieldwork relates to the administrative process, to the U-sample undercoverage, and to limitations in the O-sample fieldwork.

4.1 The U-Sample

Undercoverage of the U-sample comprises mainly "forgotten" individuals in enumerated households (e.g. babies and soldiers); out-movers (those moving after census day); dwellings missed by the enumerators; and nonrespondents. We try to minimize these problems through specific questions in the questionnaire, a non-inquiry questionnaire, and a postcard left for non-contacts. Some of the missed people are captured in the RO-sample. In addition, different control and management actions are built into the enumeration process to achieve maximal coverage of all dwellings.

Regarding overcoverage, it should be noted that the U-sample data is collected using computer assisted personal interview and that the coverage part of the questionnaire resembles a PES form. Hence, the risk of duplicates, fabrications and other erroneous enumerations is very small. Any such enumerations are expected to be handled by a system of checks and edits of the USF.

4.2 The RO-Sample

IAF life-cycle. The RO-sample is file-dependent in the sense that different files may yield different "remainders". Clearly, we would have liked the RO-sample to be defined by IAF(e). In practice, it is based on IAF(l). Therefore, the final USF-defined remainder, which is used for estimation, is somewhat different from the one used for data collection. This means that there are people who were enumerated but are not listed in area *E* in IAF(e) ("wasted" interviews) and others that should have been enumerated but were not sent to the field. The suggested remedy is to try and minimize the time-lag between IAF(e) and IAF(l), and to use advanced imputation methods for people not included in the fieldwork.

Choice of address. The IAF address is a product of an algorithm that evaluates a pool of addresses for each person. If an address receives a higher score than a pre-defined threshold, it replaces the PR address. If a person lives in E_0 and one of the potential addresses is also in E_0 but the *chosen* address is in E_1, then this person will be added to the RO-sample of E_1. To minimize this effect the reliability of non-PR sources is inspected very carefully, before a PR address is replaced by another one.

Address geocoding. The definition of the O-sample relies heavily on the geocoding process, which identifies IAF addresses with a statistical area *E*. An address in a sampled EA that is erroneously geocoded in *E* enlarges the RO-sample. Note that geocoding is generally performed at the building level and that high-quality geocoding at this level requires substantial effort. Since we are most concerned about errors that move addresses between SAs and EAs, we set the highest priority on high-quality geocoding at these levels.

Matching the USF and IAF. All unmatched people listed in the O-sample are included in the RO-sample. Therefore, missing values in the USF as well as measurement errors may decrease the number of correct matches, and thus increase the RO-sample. Although we try not to increase the RO-sample unnecessarily, we are more concerned about false matches and therefore use a conservative Type-I matching error (link non-matched records).

Fieldwork difficulties. Data collection for the RO-sample starts about seven weeks after Census Day. Eligibility and census address are verified for every person in the sample list to obtain an estimate of the overcoverage parameter λ. We classify the sample into three levels of enumeration difficulty: A- easy B- reasonable and C- hard. Typical A-level people live in households already interviewed in the U-sample. The matching between the USF and IAF highlights them and they may be found quite easily. We also believe that households in missed dwellings are easy to find. Typical B-level people are those that can be found in other parts of the country (e.g., movers). The hard core of the C-level group comprises emigrants, in particular those who emigrated many years ago. The data collection process is planned to accommodate these three types. First, enumerators will return to the U-sample EAs with a corresponding RO-sample list. It is expected that most of the type A people will be counted in this phase, and

that some information will be collected for the other two types. This phase will last about two weeks. Second, phone numbers will be attached to the rest of the list, where possible, and tried by phone. The remaining cases will undergo a thorough tracing process using a variety of sources. It is expected that 20-30% of the RO-sample are long-term emigrants. Many of these will not be traced. We are in the midst of a national project that aims to predict for each person in the PR the propensity of being an emigrant. A model based on border control information, previous censuses, family relations, and various demographic and socio-economic variables is under construction. We believe that this model will help us in solving the most difficult cases in the RO-sample.

5. LOCALITIES THAT HAVE NO ADDRESS SYSTEM

The mainstream design relies on the geocoding of the IAF records at least to a SA level: geocoding enables us to identify the IAF records that belong to a sampled EA and thus enumerate most of the O-sample list through the U-sample fieldwork. Some localities, mainly Arab towns and other small villages, have no address system. For people living in these localities, the address fields in the IAF include only the name of the locality.

We classify these localities into three categories:

- *Localities enumerated in full*. For localities that are enumerated in full in the U-sample (mainly those comprising 3-4 EAs) the locality-level remainder in the IAF can be defined, and the mainstream methodology is applied.
- *Localities with useable 1995 information.* For Arab localities that did not expand substantially since the 1995 census (i.e., with no new neighborhoods), geographical coordinates might be obtained from the 1995 census file, enhanced by appropriate updates. This option is considered for the Arab population because of its relatively low mobility, and the tendency of new couples to live with the husband's family. The 1995 census file will be linked to the IAF, using ID numbers, the geographical information from the 1995 field enumeration will be added to the IAF, and the data updated for new familial relations since 1995 (marriages and children). Once the geographical information is attached to the IAF, and the usefulness of the data is verified, such a locality can be handled by the mainstream procedure.
- *Other localities*. This category includes all other localities that have no address system.

For localities in the last category two *independent* coverage samples will be selected. For the U-sample, the area of the locality will be divided into EAs using data such as the volume of residential buildings retrieved from the Buildings Register. A sample of EAs will be selected and all eligible people enumerated. This sample provides an estimate for the undercoverage parameter p_{1+} in a similar way to the estimate suggested in (1): the estimate is the number of people who are counted in the sample and are listed in the IAF in that locality out of the total number of people who are counted in the sample. Note that for localities that are divided into SAs, IAF eligibility is determined at the *locality level* rather than at the *SA level*, as in the mainstream process; that is, since SAs cannot be identified in the IAF, a person is counted as listed in the IAF if she lives anywhere in the locality and not in a specific SA. However, separate undercoverage estimates are obtained for each SA.

For the O-sample, an independent stratified sample of people is selected. The strata are defined by demographic variables, such as age groups, that are related to the propensity of being eligible in the locality level. The overcoverage parameter λ is estimated by the number of ineligible people in the sample out of the number of eligible people in the O-sample, *corrected for IAF undercoverage*, as estimated by the U-sample. Finally, for localities that are divided into SAs, the O-sample will also estimate the population distribution by SAs. The estimates are equal to the number on eligible people in each SA out of the total number of eligible people in the locality.

6. DESIGN OF THE COVERAGE SAMPLES FOR THE FIRST EXPERIMENT

The IC paradigm was implemented for the first time in May 2002 for one town near Jerusalem. This town has grown from about 25,000 people at the end of 1995 to more than 50,000 people at the time of the experiment. About 17,000 of the newcomers populated new neighborhoods. In addition, the population is heterogeneous and includes, among

other, new immigrants from the former Soviet Union and ultra-orthodox residents. Therefore, it was expected that the experiment would provide insight into a variety of registration patterns in the administrative files.

The town comprises 12 SAs, which are divided into 247 EAs (Table 1). The requirements for the sample design are that about 50 EAs are to be selected and that the accuracy of the population estimates be about the same for age by sex subgroups within SAs. For this experiment it was also required that a minimal sample size of 3 EAs will be selected in each SA. Application of the sample allocation formula (3), requires some estimates of the population size and of the coverage parameters. Population size was estimated using IAF(s) totals. To estimate the coverage parameters, the 1995 census file was matched to the 1995 PR. Logistic regression models were built for the 1995 data to predict p_{1+}, p_{+1} and $r = p_{1+}/(p_{1+} + \lambda)$. Explanatory variables in the model included indicators for ultra-religious neighborhoods, immigration, and several age groups. Estimates for new neighborhoods with no 1995 data were projected from other areas with similar socio-demographic characteristics. Columns (2)-(5) in Table 1 show the parameter values that were used for the design.

The next step is to find the RSE level that satisfies the overall sample-size constraint, by finding the root of $f(\alpha)$ of Equation (4). Figure 1 shows the function $f(\alpha)$ and its root for various scenarios ($\boldsymbol{m}=50$). Part (a) shows the sensitivity of the root to changes of ±10% in the value of p_{1+} of column (4) in the Table, with all other parameters fixed at their Table values. The resulting RSEs are 2.1% for the reference values (middle line in the figure), 1.6% for the higher values (lower line), and 2.5% for the lower values. Hence, the sensitivity of the RSE to relatively large changes in the value of p_{1+} is small. Part (b) of the figure shows the sensitivity of the root to changes in N (column (2) in the Table) with all other parameters fixed at their Table values. It is seen that for subgroups with size of about $N/8$ (e.g. four age groups by sex) the resulting RSE is 5.8%, and for $N/16$ it increases to 8.3%.

The allocation is computed using Equation (3). Column (7) in the Table illustrates the resulting allocation for the SA estimates, using the Table parameter values. The actual allocation (column (8)) was based on further analyses and considerations, but is very similar to the previous column. One completely new neighborhood, representing itself, was added to the sample. Finally, the resulting estimates and their RSEs are displayed respectively in columns (9) and (10). In general, the RSEs comply with the design requirements. For SA #5 the RSE is much smaller than required (0.9%), which suggests that a smaller sample would have sufficed for this area.

7. CONCLUDING REMARKS

This paper is concerned with the main features of the coverage samples in the Israeli IC paradigm. Additional ideas will be examined in future experiments. For example, we intend to check the feasibility of a two-stage design, where SAs are stratified into subgroups defined by either regional councils or localities, as appropriate, and a sample of SAs is selected within these strata.

The target population comprises all people who lived in Israel for more than one year on Census Day. Hence, part of the foreign workers population and other non-Israeli people should be enumerated. The PR does not cover this population and there is no other reliable data about them. It was decided, therefore, to separate this special group from the main target population and to estimate its size directly from the U-sample data. Hence, the sampling fractions in areas with high percentages of foreign workers will be increased to facilitate reliable estimates for this population.

Another important target is to supply socio-economic information. Concerns regarding the quality of these estimates will also be incorporated into the design. Last but not least, it should be mentioned that people in institutions are fully enumerated in a separate operation. These people are deducted from the IAF for the estimation of the coverage parameters and then added back.

REFERENCES

Blum, O. (1999). "Combining Register-Based and Traditional Census Processes as a Pre-defined Strategy in Census Planning", *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*. http://www.fcsm.gov/99papers/blum2.html

Blum, O. and R. Calvo (2001). "Geospatial Data Collection and Analysis as Crucial Processes in an Integrated Census", *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*. http://www.fcsm.gov/01papers/Blum.pdf

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*.

Table 1: Design of the coverage samples for the 2002 experiment. Population size, *N(s)*, parameter values used for the design, number of EAs, *M*, suggested allocation for α=0.02 and to sample size *m=50 EAs*, actual allocation, resulting estimates and their RSEs.

(1) SA	*(2)* *N(s)*	*(3)* p_{+1}	*(4)* p_{1+}	*(5)* *r*	*(6)* *M*	*(7)* *m(0.02)*	*(8)* Actual *m*	*(9)* $\hat{N}$	*(10)* $\hat{\alpha}$
1	2670	0.85	0.80	0.75	12	4	3	2337	0.018
2	4705	0.85	0.75	0.80	21	4	4	4689	0.017
3	4411	0.85	0.85	0.75	20	3	4	3559	0.021
4	3747	0.90	0.75	0.90	19	4	4	4819	0.019
5	6323	0.50	0.80	0.95	31	6	8	5838	0.009
6	4062	0.92	0.70	0.85	20	5	4	3355	0.016
7	2155	0.90	0.75	0.90	10	3	3	2537	0.014
8	1390	0.93	0.85	0.70	5	2	3	1323	0.012
9	2456	0.91	0.70	0.90	12	4	4	2453	0.015
10	4959	0.93	0.85	0.75	24	3	3	4526	0.020
11	4531	0.80	0.70	0.95	26	6	5	16804	0.009
12	9178	0.80	0.70	0.95	47	6	7		
Total	50587				247	50	52	52240	0.005

Figure 1: Sensitivity of the root of $f(\alpha)$ to (a) changes in p_{1+} and (b) changes in N ($m=50$).