Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

Editing and Imputing Income Data in Integrated Census: lessons learned from the 2008 Census - toward the 2020 Census

Prepared by Yael Klejman, Central Bureau of Statistics, Israel

I. Introduction

1. The Israeli 2008 Population Census was an integrated census. It combined administrative data for 100% of the population with data obtained from a large sample survey (approximately 17% of the households in the country).

2. The field survey of the census served two main purposes:
- Estimating over-coverage and under-coverage of the administrative data within defined geographic units (statistical areas);
- Collecting socio-economic information unavailable from administrative sources, such as labor force characteristics, household typology, education, housing, ownership of durable goods and disability.

3. The Socio Economic file (SEF) included all individuals who answered the field survey questionnaire.

4. The data regarding income was not collected from the field survey but rather was added to the SEF after the survey was conducted. The income was taken from the following administrative files:
- Income Tax Authority file containing income of employees (gross income and months of work) and income of self-employed (gross income only);
- National Insurance Institute file containing data on allowances paid (non- work related income).

4. The use of administrative files has many advantages (among which is reducing response burden), but it was found that those files are not perfect and some editing and imputation actions were necessary.

5. This paper describes the actions that were taken in order to provide the most accurate and complete picture of the income data of the population, a brief description of the evaluation process and possible improvements planned toward the 2020 Census, with regard to the topic of income.

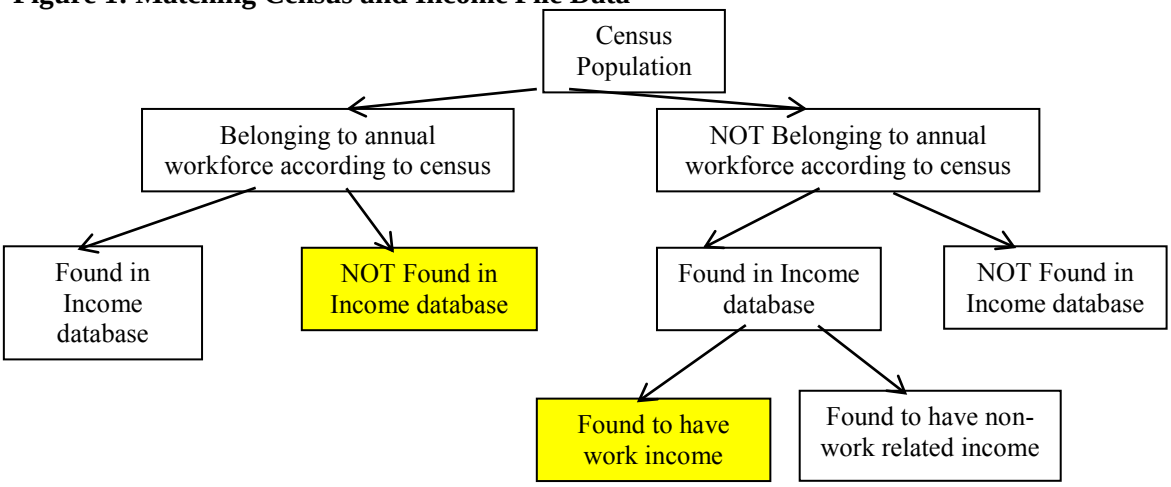
II. Imputation of income variable

A. Population definition

6. The main administrative source for the Improved Administrative File (IAF) is the Population Registry (PR). The PR contains personal records for all citizens and permanent residents of Israel and includes demographic and residential information, as well as links to records of parents, spouses and children. All registry records are identified by a unique "personal identification number" (PIN), which can be used for matching of records.

7. The following scheme presents the definition process of the population having income from work, using PINs of each individual for linking SEF records with Income Tax Authority records, in the 2008 Census:

Figure 1: Matching Census and Income File Data



8. The populations having discrepancies between their attributes are the following:
   a) Group A: Individuals who reported as belonging to the workforce, but did not have income in the administrative files;
   b) Group B: Individuals who reported as NOT belonging to the workforce, but had income in the administrative files.

B. Examination of Group A

9. Approximately 12% of the individuals, who reported to the Census as being employed, were not found in the income database for the year 2008. Of those individuals, 86.3% were employees, 12.7% self-employed, 0.9% Kibbutz or cooperative members, 0.1% unpaid family members.
It was assumed that the last two groups (kibbutz members and unpaid family members) do not report to the Tax authority due to their unique employment type. It was decided not to impute income for these groups.

10. Therefore emphasis was put on identifying the employees and self-employed individuals. It was assumed that the lack of information on these two groups is due to late or failed reporting by employers and self-employed individuals to the Tax Authority. This phenomenon appears every year at an extent of 5-7% of the employers, most of which complete the report in subsequent years, while others remain "incomplete" (due to closings of businesses, mismatching in the deductions or a real lack of reporting).
Some organizations are exceptionally tardy with their filings and it is not feasible to wait until all employers have reported. If an employee, whose employer is entirely missing from the 2008 file, was identified in the 2007 file under the same employer, then his income was imputed and indexed to the year 2008.

C. Examination of Group B

11. 18% of the individuals reporting as not belonging to the work force in the Census, were found in the administrative income database. Of these, 73% had income from salary or positive business income, while the others had non-work related income (payments received from an employer, such as pension, compensation, etc.). This non-work related income was not introduced to the SEF.

12. Some of the discrepancies were explained by irregular employment patterns or by response by proxy (family member responding to questionnaire for other family members).

13. While the exact cause of the discrepancies was not known, it was clear that this work-related income was available to the individual and the household. Therefore it was decided to include it, with a special status.

D. Income imputation for under-covered occupations

14. Military personnel and caretakers were not covered in the Tax Authority file, neither could their income pattern be traced in previous files. Income for these groups was imputed according to their income as reported in a current income survey conducted by the CBS (based on pooling the samples for years 2007-2009). The average income of the holders of these occupations in the survey was imputed to census records, by occupation group.

15. The CBS does not receive income data on army personnel for security reasons. The lack of caretakers' income data is explained by the fact that they fail to report to the Tax Authority, in case their earnings are small.

16. Caretakers: The different imputation models that were examined for this group and rejected were:

- Cold-deck: taking the individual's income from the income survey: the individuals missing from the Tax Authority File were not found in the income survey;
- Nearest neighbor from Tax Authority file: the individuals for which income was missing, had different socio-economic characteristics from the rest of the population and their income was not similar.

17. Therefore it was decided to apply a statistical imputation model based on the Income survey, sampling for the normal distribution, based on averages and standard deviation. The variables used were: extent of the job (full/partial), occupation, industry and age group. The wages of caretakers were found to have a normal distribution in the Income Survey.

18. The imputation was applied in stages:

- Average income was calculated (employee and self-employed) from the extent of the job (full or part time), occupation, industry and age group from a sample out of the income survey.
- If no fit was found by extent of the job, occupation, industry and age group, the wages were imputed by weighted average income (by work hours) per the extent of the job, occupation and industry.
- If no fit was found by occupation and industry, the average weighted income was calculated by extent of the job and occupation.
- It was important to maintain a similar distribution of income (similar variance) in imputed records vs. survey file. In order to achieve similar variance in income, for every missing record of a caretaker a record was randomly selected from the normal distribution with average income and standard deviation from income survey (by extend of the job, occupation, industry and age group).

19. Army personnel: army personnel were identified as those individuals who responded to the census question "Have you worked in the past week" with: "Was in standing army", or based on their occupation answer. No data for these individuals were found in the income file.

20. A statistical imputation was performed by averages in sub-groups, based on: extent of the job (full/partial), occupation and age group. If no fitting match was found with those variables, the weighted average was calculated (from hours worked) by extent of the job and occupation.

E. Income imputation for rest of population

21. For the rest of the population, the imputation method of the "nearest neighbor" was applied, using the Canceis program developed in Canada. The imputation was done at the individual level, taking into account all the socio-economic characteristics of the individual (as opposed to the household level, which would have provided fewer donors).

22. The following variables were used in order to identify the nearest neighbour: extent of the job (below or above 35 weekly hours), occupation (3 digits), industry (3 digits), highest education degree obtained, gender, age group, residential locality, number of children living in household per woman, marital status. Separate income imputation was conducted for residents of institutions and included "type of institution" in the variables that were used to identify the nearest neighbour.

23. Donor population: the file included individuals in yearly workforce excluding members of Kibbutz. The following were not considered as donors: caretakers and military personnel for whom the income was imputed; the highest income percentile of each occupation (2 digits).

Results of the imputation process:

Imputation type	Percent
Records with income in 2007 Income File	15.8%
Income imputation based on Income Survey for military personnel	3.3%
Income imputation based on Income Survey for caretakers	4.7%
Nearest neighbour imputation for institution residents – Canceis	2.9%
Nearest neighbour imputation– Canceis	18.3%
Records with income from Income File- not in workforce as of Census	55.0%
Records for whom income was imputed	**100.0%**

F. Status

24. All imputations were documented in an additional variable called "imputation status of income from work". This allowed the end user to decide whether and in which manner to use the data.

II. Additional editing on income variable

A. Topcoding

25. In order to avoid the possibility of discovering an individual's identity based on their exceptionally high income data, it was decided to apply a "topcoding" procedure. Several alternatives were examined, after which the following procedure was implemented:

- Detection of exceptional records was done at the locality level;
- In each locality a threshold was defined, above which all records were included in the topcoding procedure;
- For each locality in which one exceptional record was identified, the procedure was applied to at least three records (meaning a few records that did not pass the threshold were included);
- Identification of exceptional records was done by applying a threshold of interquartile range multiplied by a factor (factor 4 for urban locality, factor 3 for rural locality);
- Replacing income: the records that were identified had their income recalculated using the average of all those records. This method maintained the overall distribution and average of the income data in the locality. The procedure was documented by way of applying the relevant imputation status for those individuals.

B. Income from allowances

26. In addition to the work income, allowances were imputed based on the National Insurance Institute (NII) file. All the PIN's from the SEF, in addition to a certain number of PIN's from the Central Population Register (to prevent future identification by the NII) were sent to the NII. The NII provided data on the types of allowances paid to the individuals during the year 2008, the sums of the allowances and the period for which they were received. Data on eight different allowance types were provided:

unemployment, income support, child support, disability, survivors, child allowance, elderly support and other allowance.

27. This information was kept in a side file and was used to calculate additional variables in the SEF file (tables for individuals and households). The calculated variables were: total individual income from allowances, average individual monthly income from allowances, number of months with income from allowances, total household income from allowances (yearly and monthly average). Allowances for individuals below 17 years old were not introduced to the individual SEF table, but were added to the total household income from allowances. Codebook to clients includes a note stating that the sum of income from individuals does not always add up to total household income.

III. Evaluation of the imputation process

28. After applying the methods explained above, an evaluation of the final product was performed, based on the following:

- Comparison of the average income among the imputed records versus the average income in the income survey, by several variables: highest education degree obtained, age group, occupation: all averages came back similar.
- Comparison of basic statistics relating to income between the imputed records versus the non-imputed records: variance was maintained, average, mode and median are similar, similar positive Skewness.

29. The evaluation showed that the imputations had very satisfying results.

IV. Summary

30. Income imputations for the socio-economic file (SEF) are part of a larger activity of imputations for socio-economic variables (highest education degree, employment status, number of rooms, etc). However, income imputation has unique features, in regards to the imputation process and the data sources used. The main characteristic that differentiates income from the other socio-economic variables is its great variance. Therefore, special methods were applied in order to maintain the income distribution.

31. After careful examination, including many data simulations, the following imputation methods were used:

- Specially designed nearest neighbor method;
- Statistical methods for the unreported occupations (caretakers, military personnel).

32. The data used for the imputations were the Tax Authority File and the income data survey conducted by CBS for the years 2007-2009.

33. The conclusion reached from the evaluation process is that the carefully chosen methods of computation have proven to be reliable, by maintaining the income distribution and other statistics, which is proof that meticulous imputation methods can provide good results.

34. It is important to note that since the income was imputed from administrative files, no criteria were defined regarding "unreasonable income levels" and the income levels found were used without correction.

35. A possible improvement that is being examined for the next census is substituting the locality of residence with the socio-economic index of the locality, for the "nearest neighbour" imputation method. This will allow a larger pool of potential donors.

36. The process in 2008 was implemented on the Israeli civilian workforce and to the standing army personnel. It did not account for the income of the foreigners living in Israel. This is a challenging population we are trying to address in anticipation to the 2020 Census, since it appears that administrative

files are proving to be more reliable over time and may be able to provide a framework for addressing the foreigners. Special imputation processes will have to be developed for this population.

Bibliography

Furman, Orly, and Romanov, Dmitri (2011) *Imputation and Editing of Income from the Administrative File in the Census*. *United Nation Economic Commission for Europe*, Work Session on Statistical Data Editing, Ljubljana.

Rotenberg, Eva (2011) *Documentation of the Recording of Income for Social-Economic Data Base 2008*, Technical Paper, Central Bureau of Statistics, Israel (in Hebrew).