WP.44
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (vii): New and emerging methods

Imputation and Editing of Income from the Administrative File in the Census

Invited Paper

Prepared by Orly Furman and Dmitri Romanov, Central Bureau of Statistics, Israel

I. Introduction

1. Based on analysis of quality of the income data in the 1995 census, the Central Bureau of Statistics has decided to adopt a change in the data collection plan for the 2008 census. Whereas in previous censuses, the respondents were inquired as to their work and income, it has been decided that in the 2008 census the respondents will be inquired as to their work, while the income data will be obtained from data bases of the Tax Authorities that are available to the Central Bureau of Statistics, and imputed for each individual record in the census.

2. Initially, an overview will be provided of the examination of the income data for the 1995 census, which will be followed by a description of the consistency checks for the 2008 census and concluded by a presentation of the algorithm of income imputation and editing in the 2008 census.

II. Examination of the Income Data for the 1995 Census

A. Definitions

3. **Matched Employer-Employee Data Base -** An annual administrative data base that includes information on all employee jobs in the economy, as reported by the employers to the Tax Authorities. Each record in the data base is identified by the employee's and employer's unique identification numbers. The data base includes information concerning the income of the individual, specific tax provisions and details of the months worked during the tax year, which is the calendar year. Multiple records may exist for the same individual in this database in respect of several concurrent jobs filled during the year or the transition between jobs.

4. **Data Base of Self-Employed -** An annual administrative data base that includes all the reports of self-employed individuals to the Tax Authorities. The data base includes information concerning the income of the individual and his/her spouse. This data base does not detail the months worked by the self-employed individual during the year.

5. **Data Base of Individual Income -** A data base that was constructed on the basis of the Matched Employer-Employee and the Self-Employed data bases. This data base consolidates the information for each individual with respect to employee jobs and self-employment. Each record in this data base represents one individual, and each individual has one record.

6. **Monthly Salary per job -** The average salary per employee job for job *i* is calculated as follows: $w_i = W_i / N_i$ with W_i signifying the annual salary in job *i*, and N_i signifying the number of months worked during the same year in that job.

7. For purposes of editing and supplementing the income data that had been collected in the 1995 census, the administrative salary data base for 1995 was obtained from the National Insurance Institute, which comprises information on the annual salary of employees in all jobs that had been reported to National Insurance and details of the months worked during the year.

8. Information on salaries in the 1995 census was collected from 20% of the population that were requested to fill out an expanded questionnaire that presented a variety of questions on social-economic issues. Some of the salary data for the month of record in the 1995 census (September) have been amended or imputed during the editing of the census data as a result of missing responses to the salary question.

9. Table 1 presents the distribution by the types of amendments made during editing. The most common types of amendments performed were imputation by regression (21%) and imputation of values from the National Insurance data base (8.6%). For 69% of employees whose salary appears in the census data base, the original figure was retained.

Table 1. Amendments Made to the Salary Data in the 1995 Census

Treatment/amendment	Percentage of total
Non-amended value	69.0
Gross salary imputed by regression from net salary	21.0
Imputation of data from administrative data base	8.6
Editing of irregularities (division of income by 100/10)	1.2
Other editing	0.2
Total	100.0

B. The reported salary figure retained

10. The wording of the question concerning the salary was as follows: “What was your **gross** income (before deductions) **from your salary** in September 1995?” [highlighted in original]. However, if the individual does not know the amount of his/her gross income, it is permissible to report the net income, with the appropriate denotation.

11. Since the data collected in the census was of gross income and net income, and the net income data have been unfortunately lost over time, it may only be assumed that the income data for employees indeed reflect gross salary. To examine this assumption among the employees whose salary figure has not been amended, reconciliation was made between the reported salary and the gross/net monthly salary per job as calculated from the National Insurance data base.

12. Chart 1 presents the deviation of the reported salary from the computed gross salary in relation to the deviation from the calculated monthly salary less compulsory payments deductions (personal income tax, social security and health insurance contributions) - assuming that the reported salary is net income. On the one hand, Chart 1 supports the assumption that with regard to low and medium income (percentiles 10-50 in approximation), the reported salary could be the net salary, as its deviation from the net calculated salary is close to null. On the other hand, the divergence of the two lines to opposing sides of the 0 axis in the five highest deciles suggests that, at the top end of the income distribution, the under-reporting of income appears to be common.

13. Chart 2 indicates a tendency to round the reported salary. The extraordinary frequency of the presentation of reported salary rounded to thousands and half-thousands shows that individuals indeed tend to round their income. The rounding of the amounts may potentially affect the accuracy of the income data.

14. It has been found that the reported figure was, on average, 15.6% less than the salary recorded in the National Insurance data base. In the lower ranges of income (the first 27 percentiles in the sample), income is over-reported, while the respondents with medium to high income tend to under-report their income.

Chart 1. Deviation of the September Salary Reported in the Census from the Gross and Net Monthly Salary Per Job in the National Insurance Data Base, by Annual Salary Percentiles, as a *percentage* of gross calculated salary

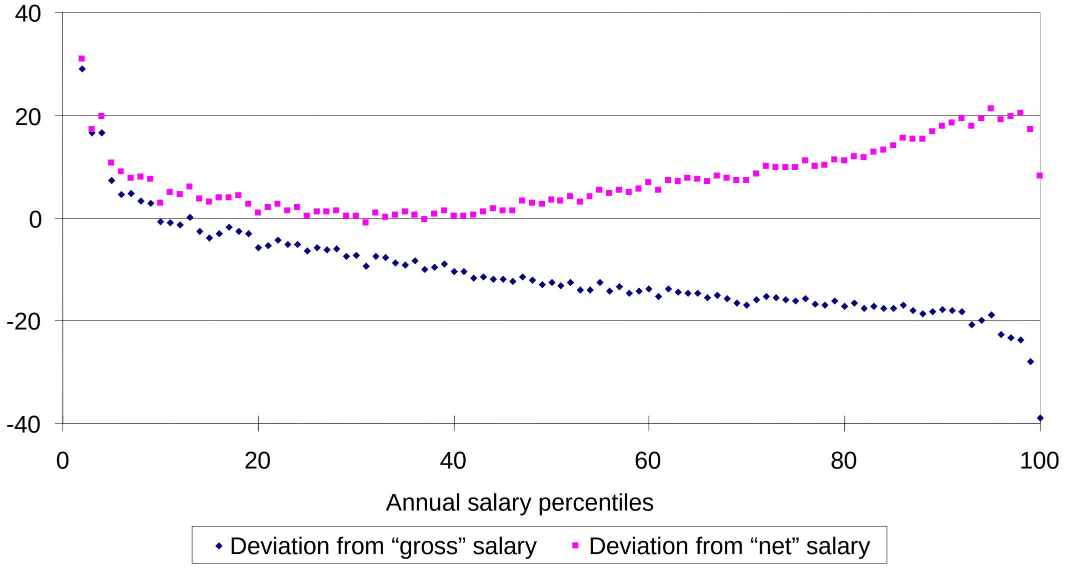
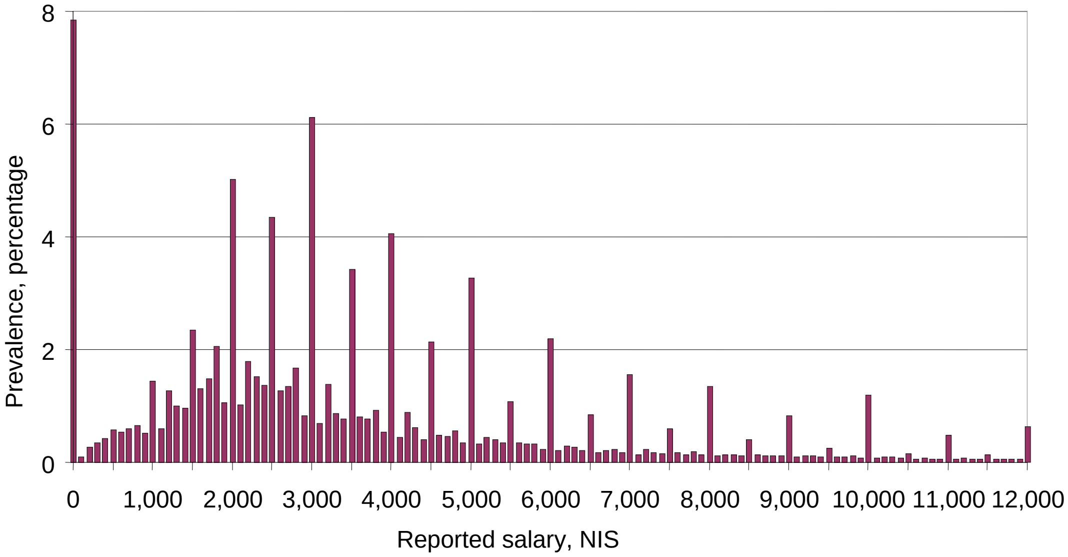


Chart 2. Distribution of September Salary Reported in the Census



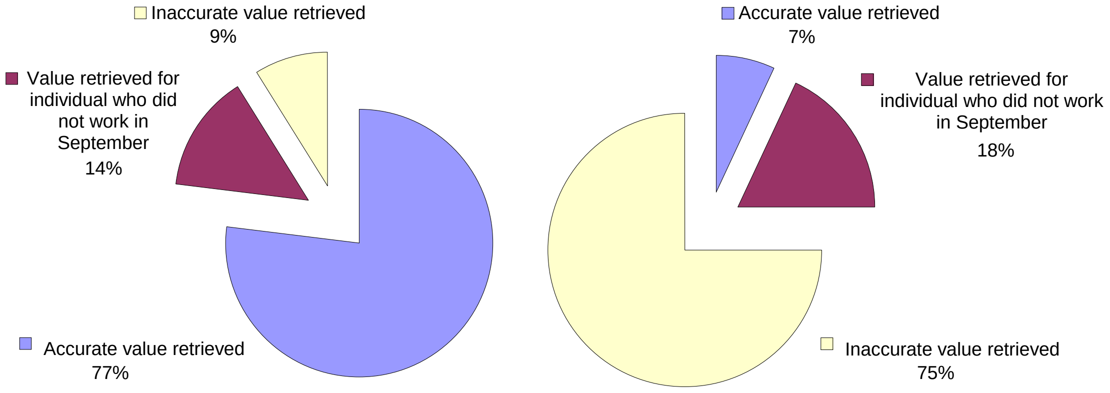
C. Imputation of Income from the National Insurance Data Base

15. In certain cases in which the salary values have been imputed out of the National Insurance data base, discrepancies were found between the figures recorded and the original figures. Discrepancies were found to be prevalent in those cases where the employee had held several jobs during the year - in 75% of these cases, the calculation of the salary for September was inaccurate. Additionally, in 15% of the cases, salary was recorded for September although the individual did not work in September (Chart 3).

D. Imputation by Regression

16. The imputation of data through regression was also inaccurate. Regression was designated to extrapolate gross salary values from net salary values, since the questionnaire allowed the indication of both values or either of them. In those cases where only the net salary value had been stated in the questionnaire, a gross salary regression was imposed on net salary in four segments, with slopes of 1.175 to 2.013. A significantly more accurate estimate could have been achieved by using a calculation by income tax function and schedules of the national insurance and health insurance contributions.

Chart 3. Distribution of Values Imputed on the Basis of the National Insurance Salary Data Base for Employees who Held One Job (Left) and More Then One Job (Right) in 1995



17. Following the identification of deficiencies in the reporting and editing of income in the 1995 census, it has been decided to fully impute the earnings income for the 2008 census out of an administrative data base.

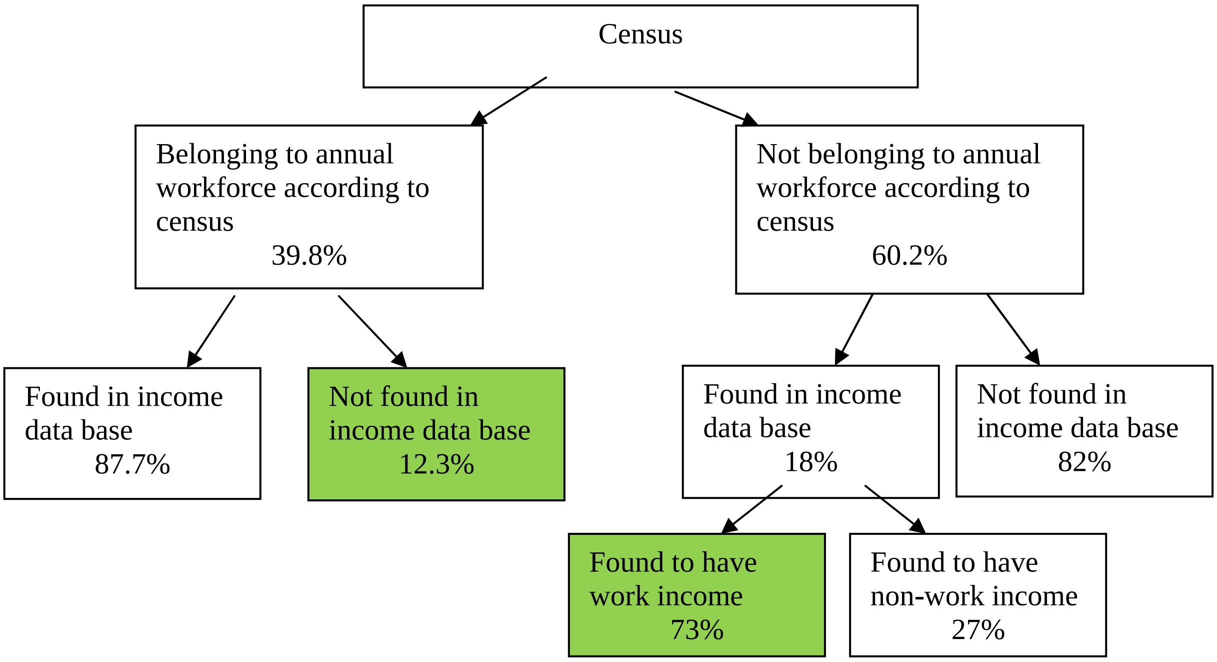
III. Imputation and Editing of Income Data in the 2008 Census

A. General

18. Unlike previous censuses, in which the respondents had been inquired as to their work and income, in the 2008 census the respondents have been inquired as to their work, while the income is to be imputed out of the administrative data bases for each individual.

19. Within the framework of the matching of the census data with the administrative data that include all employees and self-employed individuals, in 87.2% of cases employment details as reported in the census and earnings details as recorded in the administrative files fully coincided. For the remaining cases, an algorithm for the imputation of income needed to be created. Chart 4 presents the algorithm for the identification of said cases.

Chart 4. Flow of the Identification of Cases in Which the Census Data and the Administrative Data do Not Coincide


Census
Belonging to annual workforce according to census 39.8%
Not belonging to annual workforce according to census 60.2%
Found in income data base 87.7%
Not found in income data base 12.3%
Found in income data base 18%
Not found in income data base 82%
Found to have work income 73%
Found to have non-work income 27%

20. Two principal groups have been identified in which discrepancy exists:

(a) **Group A**: Individuals that were found to have work income in 2008 as per the administrative income data base, which according to the census did not belong to the annual workforce.

(b) **Group B**: Individuals that reported in the 2008 as belonging to the annual workforce, but were not found to have work income according to the administrative data bases.

B. Analysis of Group A

21. 18% of the individuals who were not classified as belonging to the annual workforce according to the census have been found in the individual income data base for 2008. Of these, 73% have income from salary or positive business income, while the others have non-work income (payments received from an employer, such as pension, severance pay etc.).

22. The record date in the census questionnaire was December 29, 2008, so that the inquiry of employment in the past 12 months coincides with the period covered in the administrative data base for said year. Accordingly, the work hypothesis in examining these cases (individuals having work income as per the administrative income data base) is that the reporting in the census of not belonging to the workforce is incorrect. Possibly, this is due to a part-time or temporary job that the respondent forgot to mention, but it is also likely that some individuals deliberately left out details of their work.

23. 67% of this group is in the primary working age-group (19 to 65). 51% of this group worked in 2008, according to the income tax data, less than half a year. This fact reinforces the hypothesis that the possible reason for the under-reporting of employment in 2008 is irregular employment over this year and the forgetting of the employment period in the census report. In order to test this hypothesis, those individuals who had been employed in December 2008 as per the income data base were examined. For these individuals, the forgetting of the job is unlikely, but is probably the result of incorrect reporting.

24. The examination of the work in December 2008, which is the record month in the 2008 census, revealed that for 43% of Group A, a record exists in the administrative income data base for that month. It should be kept in mind that self employed individuals do not have information as to work months and therefore this figure is by its nature absent. 74% of the individuals having work income in December who did not report employment in the census were found to be between the ages 19 to 65, and for two thirds of them the income data base includes information on ongoing employment in 2008, for over six months of employment. This information indicates a high probability of inaccurate reporting in the census with respect to labour market non-participation.

C. Analysis of Group B

25. 12.3% of the individuals who reported employment in the 2008 census, i.e. belonged to the annual workforce, were not found in the individual income data base for 2008. According to the census, 80% of the above worked 12 months during 2008. 84% of this group are in the principal working ages (25-65). The distribution of their work status is presented in Table 2.

Table 2: Distribution by Work Status of Individuals who Reported Employment in the Sample but Were Not Found in the Income Data Base

Work status	Distribution as reported in the census	Absent from the income data base, % of cell
Total	**100.0**	**12.3**
Employees	86.3	10.7
Self employed – not employing	8.3	16.5
Self employed – employing	4.4	12.7
Cooperative members	0.1	25.6
Kibbutz members	0.8	75.6
Unpaid family members	0.1	51.2

26. It is probable that cooperative members, kibbutz members and unpaid family members are likely to be found in the administrative income data base at relatively low rates, since these are small and unique groups in terms of the recording of their income for purposes of reporting to the Tax Authority. Therefore, the emphasis in the examination was placed on employees and self employed individuals. The work hypothesis is that the absence of information on these two groups is due to late or failed reporting by employers and self employed individuals to the Income Tax authority, which report serves as the basis for the construction of the administrative income data base. This phenomenon recurs every year at an extent of 5-7 percent of the employers, some of which complement the report in subsequent years, while others remain “incomplete” (due to the closing of businesses, mismatches in the deductions file or genuine lack of reporting).

27. Accordingly, the employer of an employee who was absent from the 2008 income data base should be examined, to check whether the employee was active in the preceding year. Therefore, these details were examined for 2007. The examination shows that more than 50% worked in 2007 and have employee jobs. 80% of these employees work for employers that did not report in 2008 but did report in 2007.

28. The census provided data regarding the economic industry and occupation, including specification of the name of the employer. An examination performed revealed that, as a rule, it is the same employer that appears in the 2007 income database. An additional examination performed among self employed individuals who reported in 2007 but not in 2008, revealed that 7% had an active tax file in 2007. Of these, almost 83% reported self employment in the census.

D. Algorithm of Income Imputation

29. Employees found with employers in 2007 that did not report in 2008. There are active businesses that in 2008 employed employees but did not filed a report to the Income Tax Authority until the date of transfer of the data base to the CBS. Individuals who reported in the census as having worked in 2008 but were not found in the individual income database for 2008 have been identified as holding a job in 2007. For other individuals who work for employers that did not report in 2008, the income found in 2007 was imputed, after being adjusted according to the nominal change between the years 2007 and 2008 in average salary, by the economic industry. The salary was imputed based on the number of work months reported by the individual in the 2008 census.

30. Self employed individuals who reported in 2007 but did not report in 2008. Similarly to the process described above with respect to employees, individuals who had reported in the census having worked in 2008 but were absent from the individual income data base for 2008, were identified in the 2007 tax files, out of which income was imputed for those self employed individuals who did not report in 2008. The business income for 2007 was imputed in 2008, after being adjusted by the nominal change between the years 2007 and 2008 in the income of self employed individuals, by the economy industry.

31. Military personnel, housekeepers and caretakers. For this group, income may not be imputed on the basis of the administrative income data base, since this group is not covered in said data source, and its income patterns cannot be traced. Income for this group was imputed according to a current income survey existing in the CBS. In this survey, the holders of these occupations were identified and the income was imputed, by occupation group, according to the average income in the survey (based on pooling the samples for years 2007-2009).

32. Individuals who reported having worked in the census that do not belong to the above-mentioned groups. Income was imputed within the estimation cells in a multi-dimensional matrix that was constructed on the basis of variables such as economic industry, occupation, age group, gender, marital status, education level. The method employed is the matching to the “closest neighbour”, i.e. for each individual in this group an individual is identified who has known income and that resembles the said array of economic and demographic characteristics, then the "neighbour’s income is imputed.

33. Table 3 summarizes the proposed imputation methods for the groups discussed above.

Table 3. Summary of Groups and the Manner of Income Imputation

Group	Income recording method	Percentage of total imputed cases	Percentage of cases reported in census
Found to have work income according to income data base but do not belong to the workforce according to the census.	Work months and salary imputed as per the income data base.	61.7	7.9
Belong to the workforce according to the census but found not having work income according to the income data base, found to be employed by employers in 2007 that did not report in 2008	The individual's salary for 2007 was imputed, adjusted for the average salary increase in the economic industry.	15.2	1.9
Belong to the workforce according to the census but found not having work income according to the income data base, found to be reporting self employed individuals in 2007 who did not report in 2008	Income was imputed for holders of active files in 2007, adjusted for the average income increase in the economic industry.	2.6	0.3
Belong to the workforce according to the census but found not having work income according to the income data base, military personnel, housekeepers, caretakers and unknown denotation of occupation	Income was imputed from the ongoing survey, according to the average income as per defined estimation cells*.	3.6	0.5
For individuals who reported having worked in the census but do not belong to the abovementioned groups	Income was imputed based on average income in the defined estimation cells**, according to the number of months worked as reported in the census.	16.9	2.1
Total		100.0	12.7

* Income Survey for 2007-2009. estimation cells have been defined according to age group, gender, occupation

** Estimation cells have been defined according to age group, gender, occupation, schooling, region.

Bibliography

Furman, Orly, and Romanov, Dmitri (2006) *Examination of Salary Data in the 1995 Census using the National Insurance Salary Data Base*, Working Paper No. 21, Central Bureau of Statistics, Israel (in Hebrew).

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

Yitzhaki, Shlomo (2007), *Pros and Cons for using Administrative records in Statistical Bureaus. United Nation Economic Commission for Europe*, Conference of European Statisticians, Geneva.