**UNITED NATIONS STATISTICAL COMMISSION and**
**ECONOMIC COMMISSION FOR EUROPE**

Working Paper No. 30
English only

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Madrid, Spain 20-22 October 2003)

Topic (iii): Data editing processes within survey processing

# THE USE OF ADMINISTRATIVE DATA IN THE EDIT AND IMPUTATION PROCESS

**Supporting Paper**

Submitted by the Central Bureau of Statistics, Israel[1]

## I. INTRODUCTION

1. Most statistical agencies have access to high quality administrative data which can be used to supplement and augment statistical processes. The cost effectiveness and the reliability of the administrative data must be carefully considered before putting the data to its intended uses. Administrative data can be used as a primary source for statistical output or as a secondary source to enhance statistical processes. The main uses of administrative data in the agency are:

- Building registers, sampling frames and censuses,
- Calculating auxiliary data and synthetic estimates for estimation procedures,
- Improving the edit and imputation phase of surveys with more complete and consistent data.

2. When using administrative data as a primary source for statistical information, such as for a business register, extensive editing and cleaning must be undertaken to improve the consistency of the data (especially if two or more sources are linked together), the timeliness of the data and the treatment of outliers. The data can then be used for direct statistical analysis, for defining parameters for sampling designs (i.e., size variables for pps sampling) or used as auxiliary data for regression and ratio estimators in the estimation phase of a survey. In addition, administrative data can be incorporated directly into a survey through the use of synthetic estimates for small area estimation.

3. Incorporating administrative data as a secondary source for improving survey processes can be very effective and will have an impact on the quality of the survey data. By linking survey data and administrative data, the underlying mechanisms of non-response can be better understood and estimation procedures can be improved. For unit non-response, the issue of individual imputation verses correction by sampling weights usually depends on the type of survey and the amount of external data that is available on the non-respondents. In the edit and imputation phase of a survey, auxiliary administrative data can be used to correct and complete the data set for erroneous and missing data, including those for the non-respondents. The complete data set can then be used as reliable covariates for more sophisticated models to impute other survey target variables based on the homogeneity of the response probabilities or of the known target variables.

---

[1] Prepared by Natalie Shlomo (natalies@cbs.gov.il)

4. This paper will focus on the above two themes. In the next section, an algorithm will be presented for incorporating administrative categorical demographic data in the edit and imputation phase of a survey based on a large scale cold-deck module for imputation before the use of stochastic imputations. In section III, specific problems for selective editing of administrative data will be presented, especially when two or more sources are linked together to produce a direct statistical source such as a business register.

## II. IMPROVING THE EDIT AND IMPUTATION PHASE OF A SURVEY

5. One of the most complete administrative sources at the Israel CBS is the National Population Register (NPR). The NPR includes demographic and geographic information for all citizens of the country, although for about 20% to 30% of the individuals the address information is outdated and erroneous. The demographic data, however, is very reliable. The use of the NPR data is particularly cost effective since every person in the file is identified by a unique identity number which is required upon birth or upon immigration, and therefore can be linked quite easily with survey data. Extensive use of the NPR data has been carried out for analyzing characteristics of non-respondents for specific surveys, such as the Family Expenditure survey, in order to improve current estimation practices (Yitzkov and Kirshai-Bibi (2003)). In the remainder of this section, we will present an algorithm for improving the edit and imputation phase of a survey using NPR data by checking inconsistencies and filling in illogical or missing values for categorical demographic data. The purpose is to reduce the amount of stochastic imputations necessary on the basic demographic data and to improve general imputation models for survey target variables by supplying complete and consistent demographic variables along with the geographic variables obtained from the survey data.

6. A survey was recently carried out in one town in Israel with about 50,000 inhabitants for the purpose of collecting demographic data. The town was mapped into enumeration areas, where each area had about 50 dwellings. A 20% area sample was selected and included 52 enumeration areas. In each enumeration area, computer-assisted personal interviews were carried out for all persons in the dwellings. The survey included 9,913 persons, thereof 9,422 persons were linked to the NPR by exact matching based on the identity numbers, sex and date of birth, and probability matching on the residuals based on names. The remainder were not linked to the NPR because they live permanently in the country without citizenship and do not have identity numbers. Some initial editing was carried out based on logic checks incorporated into the computerized questionnaire and crude mistakes were corrected in the database, such as deleting duplicates. On this survey data, we will demonstrate how the edit and imputation phase for the demographic data was carried out based on the NPR.

7. Before beginning the edit and imputation process, a list of explicit edit rules were drawn up by the subject matter specialists. For the purpose of demonstrating the algorithm, we will use the following 16 explicit edit rules on the demographic data:

$E_1$ ={“Sex” notin (male, female)}=Failure
$E_2$ ={“Year of Birth”< 1890 or “Year of Birth”> 2002}=Failure
$E_3$ ={“Year of Marriage” - “Year of Birth” <15}=Failure
$E_4$ ={abs(“Year of Birth” - “Year of Birth of Spouse” ) >25}=Failure
$E_5$ ={“Year of Birth ” – “Year of Birth of Mother” <14}=Failure
$E_6$ ={“Year of Marriage ” ^= “Year of Marriage of Spouse” }=Failure
$E_7$={“Sex ” = “Sex of Spouse”}=Failure
$E_8$ ={“Year of Birth ” > ”Year of Immigration”}=Failure
$E_9$ ={“Year of Birth ” – “Year of Birth of Father” <14}=Failure
$E_{10}$ ={“Marital Status” notin (married, single, divorced, widow)}=Failure
$E_{11}$ ={“Marital Status” in (married, divorced, widow) and “Year of Birth” >1987}=Failure

$E_{12}$ ={"Marital Status" =single and "Year of Marriage" ^=null}=Failure
$E_{13}$ ={"Marital Status" ^= "Marital Status of Spouse"}=Failure
$E_{14}$ ={"Marital Status" = married and "Year of Marriage" =null}=Failure
$E_{15}$ ={"Date of Immigration" = null and "Country of Birth" ^=Israel}=Failure
$E_{16}$ ={"Date of Immigration" ^= null and "Country of Birth" = Israel}=Failure

Each one of the above edit rules are defined by logic propositions. The logic propositions are in standard SAS programming language using the exact names of the fields as defined in the data dictionary. For example, the edit $E_5$ ={"Year of Birth" – "Year of Birth of Mother" <14}=Failure involves three logic propositions:

- yearofbirth>1890 and yearofbirth<2002
- yearofbirthofmother>1890 and yearofbirthofmother<2002
- yearofbirth-yearofbirthofmother<14

In order for the edit to fail on a particular record, all of the logic propositions for that edit have to be true on the record. All of the above edits were broken down into logic propositions. Out of the 16 edit rules, 34 logic propositions were constructed. The consistency and logic of the edit rules were extensively tested by using test data, although more sophisticated techniques for checking the edit rules will be further developed in the future.

8. The edit rules are defined in an edit matrix and for this purpose we used a standard Excel spreadsheet. The first column of the matrix includes all of the logic propositions for all of the edit rules. Each column following the first column of logic propositions represents one edit rule, where a one is placed in the cell if the logic proposition participates in the edit rule, and zero if the logic proposition does not participate in the edit rule. The number of columns in the matrix (besides the first column of logic propositions) is equal to the number of edit rules, or 16 in this case. The number of rows in the edit matrix is equal to the total number of propositions that make up the edit rules, or 34 in this case. The edit matrix was then imported into a SAS file.

9. Before incorporating the NPR data, the survey data was checked against a subset of the above edit rules. The survey data did not include year of marriage so edits: $E_3$, $E_6$, $E_{12}$ and $E_{14}$ were dropped. In order to check the edit rules automatically we developed the following algorithm according to the framework of Fellegi and Holt (1976):

- The logic propositions are transformed into Boolean logic statements and placed in a new SAS program which will be applied to the records in the dataset.

- As a result of running the SAS program containing the Boolean logic statements on the dataset, new fields are added to each record which contain the results of the logic propositions. If a particular logic proposition is true on the record, the value of one is placed in the field, and if the logic proposition is false on the record, the value of zero is placed in the field.

- The output of this SAS program is therefore a new matrix where each row is a record of the dataset and each column represents a logic proposition containing either a one if the proposition is true and a zero if the proposition is false on the record. In this case, we have a matrix of 9,422 rows representing the records in the data and 34 columns of logic propositions.

- The records matrix (9,422 records*34 propositions) is multiplied by the edit rules matrix (34 propositions*16 edit rules) resulting in a new matrix consisting of 9,422 records and 16 edit rules and each cell of the matrix contains the scalar product of the vectors making up the two original matrices. If the scalar product is equal to the total number of logic propositions for a particular edit, then the record fails that edit.

This algorithm was applied on the survey data and the results are presented in Table 1.

**Table 1: Failed edit rules for survey dataset**

| Edit Rules | Number of Records with Failed Edits | |
|---|---|---|
| | **Total** | **Percentage** |
| **Total Records Checked** | 9,422 | - |
| **Records Failing: $E_1$** | 171 | 1.81% |
| **$E_2$** | 203 | 2.15% |
| **$E_4$** | 5 | 0.05% |
| **$E_5$** | 3 | 0.03% |
| **$E_7$** | 2 | 0.02% |
| **$E_8$** | 0 | - |
| **$E_9$** | 5 | 0.05% |
| **$E_{10}$** | 0 | - |
| **$E_{11}$** | 0 | - |
| **$E_{13}$** | 7 | 0.07% |
| **$E_{15}$** | 4 | 0.04% |
| **$E_{16}$** | 26 | 0.28% |

In addition, 76 records had one edit failure, 172 records had two edit failures, and 2 records had three edit failures, not including the edit checks on year of marriage.

10. By incorporating the NPR data we can correct a priori failed edit rules by choosing the best values of the variables that assure that edit rules will not be violated. In this small survey, there were nine common variables between the survey data and the NPR data: sex; marital status; country of birth; year, month, and day of birth; year, month and day of immigration. Each one of the variables was checked to see if there is a discrepancy between the value in the survey data and the value in the NPR data. For each record in the survey data, all possible combinations of records were built from among the different values of the variables. In general, the number of record combinations depends on the number of variables in discrepancy and the number of data sources available. For this simple survey where there are only two data sources, a discrepancy in one variable on the record will cause two records to be constructed, each one having the different possible value for the variable and no changes in the other variables. With two variables in discrepancy between the two data sources, four possible records are constructed, and so on. If the survey data had a missing value, this combination was discarded. Out of the 9,422 records in the survey, 2,922 had at least one discrepancy in one of the variables that did not involve a missing value. These resulted in 11,050 different record combinations not including combinations with missing values according to the distribution in Table 2.

**Table 2: Number of records and record combinations for variables with discrepancies**

| Number of Variables with Discrepancies | Number of Records | Number of Record Combinations |
|---|---|---|
| **Total** | 2,922 | 11,050 |
| **1** | 1,785 | 3,570 |
| **2** | 694 | 2,776 |
| **3** | 348 | 2,784 |
| **4** | 74 | 1,184 |
| **5** | 19 | 608 |
| **6** | 2 | 128 |

11. The total number of record combinations to undergo edit checks is 17,550 records (6,500 single records with no discrepancies and 11,050 multiple records with discrepancies). For each one of the record combinations, a total variable field score is calculated which represents the validity and the reliability of the record combination. The total variable field score is the sum of individual field scores which are calculated for each one of the nine variables that vary on the record combination. In general, the individual field scores depend on weights that are defined by the user according to the variable and the source of the data in the record combination. If the value for a particular variable comes from a data source that is considered very reliable, more weight is given to the value as opposed to other values from less reliable sources with respect to missing data, incompleteness, timeliness, etc. In this small survey, it was decided that more weight would be given to the demographic variables on the NPR file as compared to the survey data. In addition, all variables in each data set would have the same weight. Thus, all values of variables coming from the NPR file received a weight of 0.6 and all values of variables coming from the survey data received a weight of 0.4. The individual field score for each of the variables in this case where there are only two sources of data and all variables in each source have the same weight is trivial and is equal to the weight itself. Variables with no discrepancies in the values between the NPR data and the survey data are defined an individual field score of 1. A more detailed description of individual field scores and total variable field scores is presented in Shlomo (2002).

12. The record combinations underwent the full edit checks. After selecting the records with the lowest number of edit failures and the highest total variable field score the results in Table 3 were obtained.

**Table 3: Failed edit rules for record combinations and records with highest total variable field score**

| Edit Rules | Number of Record Combinations with Failed Edits | | Number of Records with Highest Total Variable Field Score and Lowest Number of Failed Edits | |
|---|---|---|---|---|
| | Total | Percentage | Total | Percentage |
| **Total Records Checked** | 17,550 | - | 9,422 | - |
| **Records failing: $E_1$** | 0 | - | 0 | - |
| **$E_2$** | 4 | 0.02% | 0 | - |
| **$E_3$** | 4 | 0.02% | 1 | 0.01% |
| **$E_4$** | 22 | 0.13% | 4 | 0.04% |
| **$E_5$** | 5 | 0.03% | 0 | - |
| **$E_6$** | 159 | 0.91% | 52 | 0.55% |
| **$E_7$** | 0 | - | 0 | - |
| **$E_8$** | 3 | 0.02% | 0 | - |
| **$E_9$** | 13 | 0.07% | 2 | 0.02% |
| **$E_{10}$** | 0 | - | 0 | - |
| **$E_{11}$** | 0 | - | 0 | - |
| **$E_{12}$** | 43 | 0.25% | 0 | - |
| **$E_{13}$** | 67 | 0.38% | 14 | 0.15% |
| **$E_{14}$** | 3,205 | 18.26% | 1,162 | 12.33% |
| **$E_{15}$** | 49 | 0.28% | 14 | 0.15% |
| **$E_{16}$** | 615 | 3.50% | 10 | 0.11% |

For the records that had the highest total variable field score and the lowest number of failed edits, we compared the results of the subset of the edit checks that were carried out in paragraph 9. Only 44 records had one edit failure after incorporating the administrative data using the above method as compared to 250 records with at least one edit failure based on the survey data alone. This is an improvement in the number of records that have to be corrected using stochastic imputation. As for the total set of edit checks, 8,165 records had no edit failures, 1,255 records had one edit failure and 2 records had two edit failures. Table 4 presents the source of the data that was selected for building the final records of the survey for each one of the variables.

**Table 4: Source of data chosen for final records of the survey**

| Variable | No Discrepancy | Discrepancy between Survey and NPR File | | |
|---|---|---|---|---|
| | | Total | NPR Data Chosen | Survey Data Chosen |
| **Sex** | , | | | |
| **Date of Birth** | | | | |
| **Date of Immigration** | | | | |
| **Country of Birth** | | | | |
| **Marital Status** | | | | |

13. Most of the edit failures that still remained after applying the NPR data were a result of missing data. A considerable amount of records were corrected using a deterministic approach based on plausible imputation from other family members, in particular for religion, marital status, year of

marriage, date of immigration and country of birth. Additional administrative sources may also be available that could assist in the correction and imputation stage. The remainder of the records with missing values were imputed using hot-deck imputation by finding nearest neighbors on common known demographic and geographic variables. In this small survey, the few records that had inconsistent data were corrected manually. In the future, we are considering using NIM for the imputation methodology on the demographic surveys which was successfully applied to the Canadian Censuses (Bankier, (1999)).

## III. EDITING ADMINISTRATIVE DATA AS A PRIMARY STATISTICAL SOURCE

14. Israel CBS is in the process of developing a comprehensive business register where the bulk of the data will be obtained by linking two main administrative sources, the VAT file which contains data on the revenue of the businesses and the Social Security File (SSF) which contains data on the number of employee posts and wages. Up till now, business surveys were drawn from either one or the other of the files, depending on the survey. For most business surveys, the data is collected by investigating administrative sources further. All editing of survey data is carried out through selective editing indicators and those failing edit checks are checked manually for errors and anomalies. With the new business register and the reorganization of the business surveys, we are in the process of developing a comprehensive selective editing method on the linked administrative data which is the base for the business register. The linked administrative data needs to be edited and cleaned directly on the register in order to provide viable statistical information for the design of the samples and their estimation and weighting procedures. The editing and imputation will take into account the correlations between variables from different sources on the linked administrative data and more useful and optimal editing indicators will be developed. Besides the standard edit checks dealing with completeness, matching errors, duplicates, etc., the edits on the data itself will cover the following areas:

1. Missing data and time lags on the reported variables from the administrative data,
2. Identification of business entities that are outliers and donate most to the variance in the economic branch,
3. Inconsistent and illogical values across time for the main variables: revenue, employee posts, wages, average wages per employee posts and average revenue per employee post.

15. One of the problems with the VAT file is that some business entities report revenue every two months, as opposed to monthly reports. In addition, revenue may be missing for some of the months. Using the linked data set, it is usually easy to determine which business entities have missing data or time lags when reporting revenue and which business entities have zero revenue according to the monthly reports on the number of employee posts and wages from the SSF file. For those businesses with full monthly revenue data from the VAT file, the number of employee posts from the SSF file was found to be correlated well with revenue (average R-square across months of about 0.8 for most economic branches). Therefore, to improve the imputation for time lags and missing data of revenue we will incorporate the data on number of employee posts from the SSF file. For time lags, imputation classes were formed using the sum of the revenue for the two months and the number of employee posts. In each imputation class the proportion of monthly revenue out of the total two-month revenue was calculated for those businesses with full monthly data. By multiplying the two-month revenue reports of the businesses with monthly time lags by these proportions, the revenue was split into monthly data. For missing revenue data, imputation classes were formed based on the number of employee posts, and in each imputation class the ratio of monthly revenue to employee posts was calculated for those businesses with full monthly data. The missing revenue for businesses was then imputed by multiplying the relevant ratio in the imputation class with the number of employee posts of the business entity with the missing value. Since the system for the business register is being newly developed, previous revenue data from former years was not available. Therefore, we were unable to test at this stage whether using past data of the same business entities

for imputing and correcting is preferable to using other business entities in the same imputation class for the given year. Another problem with the VAT file is that the business entities that are incorporated have only one combined monthly revenue. The total revenue, however, for the individual business entities can be imputed based on the proportion of the number of employee posts in each business entity from the SSF file. It is important to note that before applying regression or ratio imputation outliers need to be treated or removed. A test data set was used to test the algorithms consisting of all reports from the VAT file and the SSF file for the year 2002 for 360 business entities from the industry of bakeries. Out of the 360 business entities in the test data, 127 businesses had time lags when reporting revenue and needed to have their revenue split between consecutive months, and 21 businesses had missing data for revenue on at least one of the monthly reports.

16. Outliers will be detected for each monthly report on the different variables: revenue, number of employee posts, wages, average wages per employee post and average revenue per employee post, according to those that contribute most to the variances in the 3-digit economic branch as expressed in their z-scores. After filling in for time lags and missing data in the test data as described in paragraph 15, 43 businesses had at least one z-score across all the variables reported over the 12 months above a threshold of 3. Out of these, 10 businesses had an outlier in at least one of the monthly reports for number of employee posts, 18 had an outlier in revenue, 12 had an outlier in wages, 8 had an outlier in average wages per employee post and 18 had an outlier in average revenue per employee post. Out of the 43 businesses with outliers, 24 businesses had an outlier in one of the variables across the months, 15 had an outlier in two variables, and 4 businesses had an outlier in 3 variables. Future work for the detection of outliers will be to examine distributional statistics (quantiles) and other more robust measures.

17. To find inconsistent and illogical values across time for the above main variables in the register, indicators will be calculated for each variable based on the ratio of the monthly report to its expected monthly report. These indicators are based on methodology developed by the Statistical Methods Unit at Israel CBS (Burstein, et. al. (1997)).

- The ratio of the value of the variable $X$ at time $t$ to time $t-1$ is calculated: $R_t = \frac{X_t}{X_{t-1}}$.
- The mean of the rations, $MR_t$, is calculated across the 3-digit industry code.
- The ratio between the true value and the expected value is: $Q_t = \frac{X_t}{X_{t-1} * MR_t}$.
- For those ratios that are over one, calculate: $S_t = Q_t - 1$ otherwise $S_t = 1 - \frac{1}{Q_t}$. This gives an equal distance from zero for $Q_i$ and $\frac{1}{Q_i}$.
- The mean and standard deviation are calculated for $S_t$ across the 3-digit industry code and the indicator for time $t$ is: $I_t = \frac{|S_t - mean(S_t)|}{std(S_t)}$.

The indicators were calculated as described above on the corrected data. For the case of revenue equal to zero at the previous time period, $t-1$, a small number will replace zero if the business entity is active, otherwise the non-active business entity will not be included in the check. All indicators with a value of 4 and over were deemed anomalous. Table 5 presents the results of the indicators for anomalous data.

**Table 5: Results of indicators for anomalous data**

| | |
|---|---|
| **Number of business entities with at least one anomaly in one of the 12 months for all variables** | 81 |
| **Thereof: Number of business entities with anomaly in 1 variable** | 39 |
| **2 variables** | 27 |
| **3+ variables** | 15 |
| **Anomaly in Number of employee posts** | 39 |
| **Revenue** | 20 |
| **Wages** | 27 |
| **Wages per employee post** | 27 |
| **Revenue per employee post** | 29 |

After filling in and imputing missing data, flagging outliers and anomalous data, we will develop a total score for the business entity based on the weighted indicators where each weight expresses the importance of the indicator. A decision rule will be applied in order to determine which business entities need to be checked manually according to the constraints on the budget, time table and editing staff. In the above simple example on the test data, if all weights for the above indicators are equal, we obtain the distribution for the total score in Table 6.

**Table 6: Number of business entities according to the total score**

| Total Score | Number of Business Entities |
|---|---|
| 0 | 253 |
| 0.1 | 46 |
| 0.2 | 31 |
| 0.3 | 25 |
| 0.4 and over | 5 |

## IV. DISCUSSION

18. We have shown in this paper the uses of administrative data in the edit and imputation process, both when the administrative data is used as a direct statistical source of data and when it is used to enhance survey processing. Incorporating administrative data improves the quality of the statistical output of the agency on condition that the data itself is of good quality and its limitations are known and can be corrected prior to its use.

## V. REFERENCES


Bankier, M. (1999), "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses", UN/ECE Work Session on Statistical Data Editing, Rome, Italy, June 1999, www.unece.org/stats/documents/1999/06/sde/24.e.pdf .

Burstein, A. and Gurevitz, G. (1997), "Integrated Logic Checks for the Social Security File", Report, Israel Central Bureau of Statistics, March 1997 (*in hebrew*).

Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation", *Journal of the American Statistical Association*, 71, 17-35.

Shlomo, N. (2002), "Smart Editing of Administrative Categorical Data", UN/ECE Work Session on Statistical Data Editing, Helsinki, Finland, May 2002, www.unece.org/stats/documents/2002/05/sde/21.e.pdf .

Yitzkov, T. and Kirshai-Bibi, N. (2003), "Demographic Characteristics of Non-Respondents to the Family Expenditure Survey", Report, Israel Central Bureau of Statistics, May 2003 (*in hebrew*).