STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 21
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Helsinki, Finland, 27-29 May 2002)

Topic (iii): Editing of Administrative Data

"SMART EDITING" OF ADMINISTRATIVE CATEGORICAL DATA

Invited Paper

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

I. INTRODUCTION

1. With the advent of high-quality administrative lists and powerful computers, more and more statistical agencies are relying on administrative data for improving and integrating their statistical output. Indeed, the use of administrative sources have brought about many changes in official statistics, in particular:

- the linking of administrative files to obtain a wealth of information for constructing frames and registers;
- improved quality control in the editing processes;
- model-assisted weighting schemes based on administrative and register data;
- small area estimation based on incorporating sample and administrative data.

In addition, the growing costs of conducting censuses, along with higher and differential non-response rates, have forced statistical agencies to look at alternatives to obtain census data through administrative files.

2. The Central Bureau of Statistics of Israel has begun the planning and implementation of an Integrated Census which will be based on administrative files and large scale surveys. The census will be carried out in 2006, with a small pilot in May 2002 and a rehearsal census for demographic data in 2004. To ensure maximum coverage of the census population, many different sources of administrative data will be used. Rare and hard-to-count populations not usually covered in administrative sources will be enumerated fully, such as institutionalized persons, nomadic tribes, etc. In addition to the expected under-coverage, the main challenge in building a census of this nature, as opposed to a conventional census, is correcting for the large over-coverage and out-dated information typically found in administrative files. To correct the administrative files, a Correction Survey will be carried out. The survey will be composed of a large area sample drawn and enumerated from among those linked to the GIS system. For those with poor address information not linked to a sampling area through the GIS system, a supplementary sample will be drawn and enumerated. All persons and households captured in the area sample will be investigated and linked to the administrative lists. Persons in the administrative

---

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

lists not found in the field will then be traced and investigated as to whether they belong to the census population and if so, where they are located. Population estimates will be calculated based on the dual system estimation procedure after linking the survey data with the administrative file and correcting for the over-coverage and out-dated information. An Evaluation Survey will also be carried out after the process to examine the robustness of the methods and models used.

3. One of the new projects being researched and developed for this new census methodology is a system for the logic checks and edits of the administrative sources, both before and after integrating the survey data. Starting with the rehearsal census for demographic data in 2004, a system is being developed for the checks and edits of administrative categorical data, and will be extended to include additional socio-economic data for the full Integrated Census in 2006. This paper will describe the methodology that will be used for building the census file: preparing and linking the administrative files, choosing the best values for the fields that guarantee the logic checks and edits, integrating the survey data with the administrative data and imputing for individuals and households that do not pass the edit constraints.

II. PREPARING AND LINKING THE ADMINISTRATIVE FILES

4. The administrative files that are available will be chosen to ensure the best coverage of the population and the necessary updated information, in particular the current addresses. One of the advantages in developing a census of this type in countries such as in Israel is the existence of a National Population Register where every person by law has to have a unique identity number upon birth or upon immigration. The possible administrative sources for the Integrated Census besides the National Population Register are the Property Tax Files, Electric and Phone Company Files, Postal Addresses, Social Security Registers, School Enrollment Registers, Border Control Files for incoming tourists staying for long periods and outgoing citizens away for long periods, etc.

5. Each administrative file chosen to construct the final Integrated Administrative File will undergo soft-editing, harmonization and standardized coding, and parsing of names and addresses. Soft-editing will include checking the validity of the identity numbers based on the number of digits, the ordering and the check digit. Duplicates and out-of-scope populations will be removed, including those who have emigrated or died. In the National Population Register, the identity numbers of parents and spouses are given for most of the records in the file, so administrative family units can be generated. For those without identity numbers of family members, the algorithm includes joining individuals with the same last name living at the same address. Also, married couples with different address information are joined and placed in one of the addresses most likely to be the true address. To obtain the relationships in the administrative household, a full matrix is calculated so that every member of the household has his own definition of the relationships to the other members of the household. This greatly aids in the edit checks at the household level.

6. The linking of the files is based on exact matching by identity numbers, sex and date of birth. Subsequent passes on the residuals of the matching process are based on probabilistic matching using names, sex and date of birth. The matching weights are calculated by string comparator measures, taking into account variations in the spelling of names, similar letters, and the distance between similar letters (Winkler, 1990). The matching passes are carried out under different blocking criteria, and the final residuals are examined manually. Population groups in one file that might not be represented in other files are added to the matched files.

III. BUILDING THE INTEGRATED ADMINISTRATIVE FILE

7. The linked file containing all of the data from the administrative files chosen to build the Integrated Administrative File (IAF) will undergo checks and edits as would any other file containing statistical information. In this case, there could be several possible choices of values for each field to be included

in the IAF. Naturally we want to choose the values of the fields that will satisfy the checks and edits of the census, both on an individual level and on the household level. Basically, what this means is implementing a large scale deterministic or cold-deck imputation module, where the choice of the best possible value for each field is based on passing the edit constraints and on obtaining the most reliable statistical value.

To explain the algorithm developed for this cold-deck module, consider the following example of records from two different administrative sources relating to the same administrative household:

Table 1: Example of Individuals in an Administrative Household from Two Files

File	Relation	Sex	Date of Birth	Marital Status	Date of Marriage	Date of Immigration	Country of Birth
A	head	male	19601115	married	19820416	19920820	USA
	spouse	female	19821014	married	19820416	19920820	USA
	son	male	19870214	single	-	19820110	Israel
B	head	male	19601115	married	19620416	19920000	USA
	spouse	female	19621014	married	19820000	19920820	USA
	son	male	Blank	single	-	19920820	USA
	mother	female	19421211	widow	-	19820110	Poland

With seven fields to consider for each individual, the head of the household has four possible combinations of records since both "Date of Marriage" and "Date of Immigration" have different values in the two files. The spouse has four possible record combinations, the son has eight possible record combinations and the mother one record combination. All of these individuals make up 128 possible combinations of households.

8. Starting with the record combinations of the individuals, each possible combination will receive a total field score which will represent the source of the administrative file, and the agreement pattern of the possible categories for each field. The total field score will be calculated as the sum of the single field scores. The single field scores will be calculated as follows:

8.1 Based on empirical and subjective considerations, each field i in the n linked files will receive an adjusted weight, $w_{ij}\ (j = 1..n), \quad \sum_j w_{ij} = n$. The weights represent the accuracy and the validity of field i in each of the files. For a given field, files that are more accurate with better quality will receive a higher weight. For example, files with full address information will receive a higher weight for that field than files with only partial information. The weights can also be based on empirical evaluations. For example, the proportion of blanks and invalid values in a field can be used to calculate a weight for each file, which is then adjusted to meet the constraint that the sum of the weights equal the total number of files defined for the given field. Blanks and invalid values that occur more often for a field in a particular file will result in a lower weight for that file. Files that have similar characteristics with respect to quality, validity and other considerations for a given field i will have equal weights, $w_{ij} = 1$ for all j. This might occur if some of the administrative files are “fed” off another administrative file for a particular field. For example, the National Population Register often provides data to other administrative files.

In the above example, "Date of Birth" has missing values in file B but in general the values are of high quality. The proportion of missing and invalid values in file B is 15%, so the adjusted weight will be 0.9 for file B and 1.1 for file A. Both of the fields "Date of Immigration" and "Date of Marriage" have missing values and partial missing values in file B and are of lesser quality, so the weights will be adjusted downwards. After subjective and empirical considerations, the weights chosen for "Date of Immigration" are 1.2 and 0.8 for files A and B, respectively. The

weights for "Date of Marriage" are 1.4 and 0.6 for files A and B, respectively. The other fields in the two sources were evaluated to have the same quality, so both files A and B will receive weights of 1.0 for each field.

8.2 Let $R_i = (R_{1i} ... R_{ni})$ be the profile of the categories of field i from the n linked files, and let f_{i_c} be the weighted frequency of category c in profile R_i, $f_{i_c} = \sum_{j=c} w_{ij}$. If all of the weights are equal to one, f_{i_c} will be the number of files that have category c in field i. The probability of category c being the correct value for field i given the profile R_i is $\frac{f_{i_c}}{n}$. If there is total agreement in the categories for field i in all of the files, and therefore the particular field does not cause different combinations of records, $f_{i_c} = n$ and the probability will be one. The probability defines the single field score. If the value of the field is missing or blank, the single field score will be zero.

In the above example, there are disagreements in the fields "Date of Marriage" and "Date of Immigration" for the head of the household. The single field score for "Date of Marriage" is $\frac{1.4}{2} = 0.7$ for file A and $\frac{0.6}{2} = 0.3$ for file B and the single field score for "Date of Immigration" is $\frac{1.2}{2} = 0.6$ for file A and $\frac{0.8}{2} = 0.4$ for file B. The other single field scores will all be one since the categories in the fields are all in agreement between the two files.

The initial total field scores for each of the record combinations of the head of the household before the logic checks and edits are the sum of the single field scores:

Table 2: Initial Total Field Scores for Record Combinations of the Head of Household

Relation	Sex	Date of Birth	Marital Status	Date of Marriage	Date of Immigration	Country of Birth	Initial Total Field Score
head	male	19601115	married	19820416	19920820	USA	5+.7+.6=6.3
head	male	19601115	married	19820416	19920000	USA	5+.7+.4=6.1
head	male	19601115	married	19620416	19920820	USA	5+.3+.6=5.9
head	male	19601115	married	19620416	19920000	USA	5+.3+.4=5.7

The methodology is the same for three or more files. For example, consider three administrative files with a discrepancy in one of the fields. Let two of the files agree on a category and the third file have a different category. This will result in two possible record combinations, so the single field scores will be $\frac{2}{3}$ for the category in agreement and $\frac{1}{3}$ for the category in disagreement when the files are all rated the same. If the files have different weights, for example 0.5 , 0.7, 1.8, respectively, the single field scores will now be $\frac{1.2}{3} = 0.4$ for the category in agreement and $\frac{1.8}{3} = 0.6$ for the category in disagreement.

IV. LOGIC CHECKS AND EDITS OF THE INTEGRATED ADMINISTRATIVE FILE

9. Each of the different record combinations of the individuals and the households will undergo the logic checks and edits. Since we consider here demographic data, the edit constraints are simple to convert to vectors of ones and zeros that can easily be compared to the record combinations (Fellegi and Holt, 1976). For example, the vector for the edit {"Date of Marriage"-"Date of Birth"<15}=F will have a one in the category "less than 15" and a zero in the category "greater than 15". Similarly, we can define the record combination that needs to be examined as a vector with a one placed in the category defined by the actual difference between the two fields and zero in the other. If the scalar product of the record combination with the edit is equal to one, the record combination will fail the edit constraint. These vectors can be extended to include all individuals in the household, so that more complex edit constraints may be examined to check the consistency within a household.

10. Each record combination of the individuals will be examined against the edit constraints that check the consistency of the data within the record. In the above example for the head of the household in Table 2, the third and fourth record combination fail the edit constraint {"Date of Marriage"-"Date of Birth"<15}=F. The record combinations that do not pass the edit constraints will automatically receive zero for the final total field score.

Examining the eight possible record combinations of the son, where the single field score for "Date of Birth" is $\frac{1.1}{2} = 0.55$ for file A and 0 for file B because of the missing value:

Table 3: Final Total Field Scores for Record Combinations of the Son

Relation	Sex	Date of Birth	Marital Status	Date of Marriage	Date of Immigration	Country of Birth	Initial Total Field Score	Final Total Field Score
Son	male	19870214	single	-	19820110	Israel	1+1+.55+1+1+.6+.5=5.65	0
Son	male	19870214	single	-	19820110	USA	1+1+.55+1+1+.6+.5=5.65	0
Son	male	19870214	single	-	19920820	Israel	1+1+.55+1+1+.4+.5=5.45	0
Son	male	19870214	single	-	19920820	USA	1+1+.55+1+1+.4+.5=5.45	5.45
Son	male	blank	single	-	19820110	Israel	1+1+0+1+1+.6+.5=5.1	0
Son	male	blank	single	-	19820110	USA	1+1+0+1+1+.6+.5=5.1	5.1
Son	male	blank	single	-	19920820	Israel	1+1+0+1+1+.4+.5=4.9	0
Son	male	blank	single	-	19920820	USA	1+1+0+1+1+.4+.5=4.9	4.9

The first two record combinations fail the edit constraint: {"Date of Immigration"-"Date of Birth"<0}=F, so the final total field score will be zero. The first, third, fifth and seventh combinations fail the edit constraint: {"Date of Immigration">0 and "Country of Birth"="Israel"}=F. Thus, only three record combinations pass the edit constraints and receive a positive final total field score. If no combinations pass the edit constraints, the combination with the highest total field score prior to the edit checks will be flagged and will be used in the rest of the analysis.

11. For each individual, the record combination with the highest final total field score will enter the household edit checks. If this household passes the edit checks, the process is completed and those records will enter the IAF. The household edit constraints include checks for examining the relationships in the household, the ages of the family members, and other demographic variables. By taking advantage of the full matrix defining all of the relationships in the household separately for each individual, checking the consistency of the ages and other demographic details in the household will be straightforward. Thus, the edit constraint defined as {"Date of Birth of parent"-"Date of Birth of child"<14}=F can be compared to all of the parents and their children in the household regardless of

their position in the household. In the above example, the record combinations of the individuals with the highest positive final total field scores make up the following household:

Table 4: Household of the Individuals with the Highest Final Total Field Score

Relation	Sex	Date of Birth	Marital Status	Date of Marriage	Date of Immigration	Country of Birth	Final Total Field Score
head	male	19601115	married	19820416	19920820	USA	6.3
spouse	female	19621015	married	19820416	19920820	USA	6.25
son	male	19870214	single	-	19920820	USA	5.45
mother	female	19421211	widow	-	19820110	Poland	7

This household passes all of the edit constraints on the household level and thus will enter the IAF. Households that do not pass all of the edit constraints on the basis of the record combinations of the individuals with the highest final total field scores will be further examined. From among the possible record combinations, the record combination of the individual with the next highest final total field score will enter the household and replace its record combination that was checked at an earlier stage. This results in a new combination of the household, which will then be checked against the edit constraints. By sequentially replacing record combinations of individuals with the next highest positive final total field score, all of the household combinations will be checked until one is found that passes the edit constraints.

If no combinations of individuals or households exist that ensures the logic checks and edits, the following cases are flagged for follow-up:

- All of the individual record combinations have at least one combination with positive final total field scores, but together no household combination exists that passes the household edit constraints.

- At least one of the individual record combinations has missing values or did not pass the edit constraints on the individual level, but a household was found that passed the edit constraints.

- No household combination was found that passed the edit constraints and at least one of the individual record combinations has missing values or did not pass the edit constraints on the individual level.

V. INTEGRATING THE SURVEY DATA

12. A Correction Survey for the Integrated Census will cover about 10% - 20% of the IAF. The data from the survey will undergo soft-editing and checks for gross inconsistencies. The survey data will then be linked to the IAF to allow for the dual system estimation procedure. To build the final Integrated File (IF) based on the linked survey data with the IAF, the same procedure as described above will again be implemented for choosing the best value for each field. The sample will be considered as another source to take into consideration along with the other administrative files. However, the adjusted weight for the survey data may be higher than that of the administrative files to reflect higher credibility. The highest scoring records that pass both the individual and household edit constraints will enter the final IF, as well as the records from the IAF checked in the earlier stage that weren't included in the Correction Survey. Households and individuals that do not pass the edit constraints at the different stages of the process, both before and after integrating the survey data with the IAF, will undergo imputation.

13. Only at this final stage of building the IF based on both the sample and administrative data, will the imputation process begin. Using the NIM Methodology first implemented for the Canadian Census in 1996 (Bankier, 1999), households and individuals failing the edit constraints will undergo a hot-deck imputation procedure based on potential donors that passed all of the edit constraints. Potential donor households are chosen for a failed household having the same characteristics with respect to the age and sex distribution, immigration and marital status, geographical location and other demographic variables. For each of the potential donors, fields are identified that differ from the failed household. By examining which donors ensure the minimum change pattern in order for the failed household to pass all of the edit constraints, a donor is chosen randomly for the imputation.

VI. CONCLUDING REMARKS

14. Because of the large cold-deck module that ensures that individuals and households in the Integrated File pass a priori all of the edit constraints, the scope of actual hot-deck imputations needed at the final stage will be limited. Therefore, the hot-deck imputation module will be simplified as compared to the NIM methodology and more emphasis will be placed on the building of the integrated files using the cold-deck module. This is the main advantage in editing administrative data as compared to a conventional survey or census data.

VII. REFERENCES

Bankier, M. (1999), "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses", U.N. Economic Commission for Europe Work Session on Statistical Data Editing, Rome, Italy, June1999, www.unece.org/stats/documents/1999/06/sde/24.e.pdf .
.

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

Winkler, W. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", *Proceedings of the Survey Research Methods Section, ASA* ,354-359.