CRP.1
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

THE USE OF ADMINISTRATIVE DATA IN THE EDIT AND IMPUTATION PROCESS

Submitted by Natalie Shlomo, Central Bureau of Statistics, Israel

Abstract: In recent years, Statistical Agencies are putting more effort into incorporating administrative data into their data processing either as separate statistical databases (registers, censuses) or for enhancing and augmenting survey data. The motivation is to reduce both the response burden on statistical units (persons, businesses) and the financial burden by collecting available statistical data at much less expense. A methodology is needed for the edit and imputation procedures of statistical data based on administrative data which will ensure consistent, accurate and high quality records. This paper will demonstrate the methodology on the Integrated Census that is being developed at the Israel Central Bureau of Statistics. The Integrated Census combines administrative based Census data with large scale Coverage Surveys for adjusting population estimates. The edit and imputation processes when constructing the Census Integrated Administrative File (IAF) and for correcting survey data obtained from the Coverage Surveys will be demonstrated.

I. INTRODUCTION

1. Many Statistical Agencies have access to high quality administrative data which can be used to supplement, augment and even replace survey data. Administrative data can be used as both a primary and direct source for statistical data or as a secondary source to enhance and improve the quality of survey data. The main uses of administrative data in the statistical processes are:

- Building registers, sampling frames and censuses,
- Defining parameters for designing samples and size variables for drawing the samples,
- Auxiliary data for calculating sample weights based on ratio or regression estimators,
- Covariates for modeling synthetic estimates that are used for small area estimation,
- Improving the edit and imputation phase of survey data by enabling more complete and consistent data for error localization and cold-deck imputation,
- Identifying non-respondents and developing better capture techniques and improved imputation models for both unit and item non-response.

Before incorporating administrative data into the statistical processes, it must be carefully analyzed and assessed so that bias will not be introduced into the statistical data. In particular, the coverage, timeliness and reliability of the administrative data need to be evaluated and quantified so that appropriate adjustments can be made if necessary. The cost effectiveness and the impact on the quality of capturing, editing and preparing administrative data for use as statistical data needs to be weighed against the cost for collecting the statistical data directly.

2. This paper will focus on edit and imputation processes for obtaining high quality statistical data based on administrative data. The first section of the paper demonstrates how edit and imputation procedures can be incorporated directly into the process for combining multiple sources of data (survey data, administrative data). Inconsistencies between data sources may occur because of the different quality and accuracy of the sources, raising the possibility of obtaining conflicting values for common variables. A method is demonstrated which chooses the optimum values for the variables which ensure high quality, consistent, and logical final records. The second section of the paper demonstrates how administrative data can be used in the edit and imputation processess for enhancing and improving survey data. This is possible when administrative data can be directly linked back to the respondents and non-respondents of a survey through a record linkage procedure. By filling in accurate demographic data from administrative sources, better methods for modeling and imputing both item and unit non-response can be developed and applied.

3. To illustrate the use of administrative data for constructing statistical databases and for improving edit and imputation procedures on survey data, we will use as an example the Integrated Census under development at the Israel Central Bureau of Statistics (CBS). This census is based on an Integrated Administrative File (IAF) which is constructed by linking multiple sources of administrative files. The IAF serves as the core administrative census file from which large scale Coverage Surveys are drawn. As in post-enumeration surveys, the Coverage Surveys are used to estimate under-coverage (those that belong to the census target population in the specified area but are not listed in the IAF) using a standard dual system estimation procedure. The main challenge, however, when conducting an administrative based census is the over-coverage due to out-dated address information (those that are listed in the IAF but not found in the specified area because they have moved or emigrated out of the country). Therefore, a second stage of the Coverage Surveys is used to estimate the amount of over-coverage which is then incorporated into the dual system estimation model. Based on the two stages of the Coverage Surveys, each person in the final IAF receives a coverage weight representing the net coverage for a particular area. The error localization of the survey data obtained from the Coverage Surveys is determined by the linked administrative data at the unit level. Cold deck imputation is carried out for the basic demographic variables: age, sex, marital status, religion, country of birth, etc. For other survey target variables, more sophisticated modeling can be carried out based on the complete demographic and geographic variables.

II. COMBINING MULTIPLE SOURCES OF ADMINISTRATIVE DATA

4. The Integrated Census IAF file is based on multiple administrative files that ensure the best coverage of the population and the most updated information, in particular address information. In Israel, there exists a National Population Register (NPR) and every person by law has to have a unique identity number upon birth or upon immigration. This is the most accurate source of administrative data for demographic data, although the register suffers from coverage problems because of emigrants that have left the country, non-citizens living permanently in the country, and persons not living in the addresses where they are registered. Other administrative sources for the construction of the IAF are Property Tax Files, Postal Address Changes, School Enrollment Registers, and Border Control Files for incoming tourists staying for long periods and outgoing citizens away for long periods.

5. Each administrative file chosen for the construction of the IAF undergoes soft-editing, harmonization and standardized coding, and parsing of names and addresses. Soft-editing includes 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 are removed, including those who have emigrated or died. In the NPR, 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 for constructing households 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 a vector containing the definitions of the

relationships to other members of the household. This greatly aids in the edit checks at the household level.

6. The linkage of the administrative files is based on exact matching by identity numbers, sex and date of birth. Subsequent passes on the residuals from the exact matching process are based on probabilistic matching using first and last names, sex and date of birth. The matching passes are carried out under different blocking criteria. The matching weights for names are calculated by string comparator measures, taking into account variations in spelling, similar letters, and the distance between similar letters (Winkler, 1990 and Yitzkov and Azaria, 2003). The matching weight for date of birth is based on a scaled exponential function of the difference between the years of birth from the two files being linked. In some cases, depending on the file being linked, there may be other matching variables involved. Population groups that appear in one file and are not represented in other files are added to the matched files in order to obtain maximum coverage of the population.

7. The linked administrative file containing all of the data from the multiple administrative sources needs to undergo logic checks and edits as would any other file containing statistical data. In this case, however, there could be several possible choices of values for each of the variables to be included in the IAF and we need to choose the values that satisfy the logic checks and edits of the Integrated Census, both on an individual level and on the household level. Basically, this means that we are implementing a large scale deterministic or cold-deck imputation module, where the choice of the best value for each of the variables is based on a priori passing edit constraints and on obtaining the value of highest quality. To explain the following algorithm for this cold-deck module, we will consider the example of records as shown in Table 1. The table contains records of individuals in the same administrative household from two different administrative files.

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

There are seven variables in common between the two administrative files. The head of the household has discrepancies in both "Date of Marriage" and "Date of Immigration" between the two files and therefore has four (2^2) possible combinations of records. The spouse also has four possible record combinations with both “Date of Birth” and “Date of Marriage” in discrepancy. 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 be defined a **total field score** which will quantify the probability that the record is correct**.** The total field score is calculated as the sum of **single field scores**. The single field score is calculated separately for each variable in the record combination and is based on the number of administrative files that agree with the value of the variable as well as the quality of each file. The single field scores are calculated as follows:

8.1 Let $i = 1,..,I$ be the number of administrative files for a variable j ($j = 1,..,J$). In the example in Table 1, the number of files I is 2 and the number of variables J is 7. Based on

empirical and subjective considerations, we define a weight w_{ij} for each variable j in file i. The weights must be adjusted to meet the constraint that the sum of the weights across the different files for a given variable equals the total number of files contributing to the variable: $\sum_i w_{ij} = I$. The weight w_{ij} represents the accuracy and validity of variable j on file i. Files that are more accurate and of higher quality will receive higher weights. For example, files with full address information will receive a higher weight for the address field than files with only partial address information. The proportion of blanks and invalid values in a variable can be used to calibrate a weight for each file. Files with high levels of erroneous or missing values result in lower weights. Files with similar characteristics with respect to quality and validity for a given variable j will have equal weights, $w_{ij} = 1$ for all files i. This occurs when some of the administrative files are "fed" off another administrative file for a particular variable. For example, the NPR often provides data to other administrative files and therefore there is equal quality between them.

In the example in Table 1, "Date of Birth" has missing values in file B but in general the values are of high quality. Assuming that the proportion of missing and invalid values in file B is 15%, the adjusted weight is 0.9 for file B and 1.1 for file A. Both of the variables "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 are 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 variables in the two data sources were evaluated to have the same quality, so both files A and B each receive a weight of 1.0 for each remaining variable: relation to head, sex, marital status, and country of birth.

8.2 Let $C_j = (C_{j1},...,C_{jI})$ be the profile of the values of variable j from the I linked files having weights $(w_{j1},...,w_{jI})$. Let f_{j_c} be the weighted frequency of value c in profile C_j: $f_{j_c} = \sum_{i \mid C_{ji}=c} w_{ij}$. If all of the weights are equal to one, f_{j_c} will be the number of files that have value c in variable j. If the weights are not equal between the data sources, f_{j_c} will be the sum of the weights for those files that have value c in variable j. The probability of value c being the correct value for variable j given the profile C_j is $\frac{f_{j_c}}{I}$. If there is total agreement in the categories for variable j in all of the files, and therefore the particular variable does not cause different combinations of records, then $f_{j_c} = 1$ and the probability will be one. This probability defines the **single field score**. If the value of the variable is missing or blank, we define the single field score as zero.

In the example in Table 1 and recalling the weights that were given for each variable in each file (see paragraph 8.1), there are disagreements in the variables "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 are all equal to one since the categories in the fields are in agreement between the two files.

The initial total field score is the sum of the single field scores. In this example, for each of the record combinations of the head of the household, we obtain the results in Table 2 before the logic checks and edits of the records.

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 variables. Let two of the files agree on a category and the third file have a different category. This results in two possible record combinations, so the single field scores for equally weighted files will be $\frac{2}{3}$ for the category in agreement and $\frac{1}{3}$ for the category in disagreement. 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.

9. Each of the different record combinations of the individuals and the households undergo the logic checks and edits of the Integrated Census. Section III, paragraphs 12-13 describe the exact method for carrying out the logic checks and edits according to the algorithm in Fellegi and Holt (1976). In the example in Table 1 for the head of the household shown in Table 2, the third and fourth record combination fail the edit constraint {"Year of Marriage"-"Year of Birth"<15}=F. The record combinations that do not pass the edit constraints automatically receive a zero for the final total field score.

Table 3 presents the eight possible record combinations for 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: {"Year of Birth">"Year of Immigration"}=F, so the final total field score is zero. The first, third, fifth and seventh combinations fail the edit constraint: {"Date of Immigration" ^= null 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 is flagged and will go through the imputation process for each individual.

10. The record combination for each individual with the highest final total field score enters the household edit checks. If this household passes the edit checks, the process is complete and those records enter the final 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 is straightforward. Thus, the edit constraint defined as {"Year of Birth of Mother"-"Year of Birth of Child"<14}=F can be compared to all of the mothers and their children in the household regardless of their position in the household and their age. In our continuing example, the record combinations of the individuals with the highest positive final total field scores make up the following household as shown in Table 4:

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 these records enter the final IAF.

11. 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 are further examined. From among the possible record combinations, the record combination of the individual with the next highest final total field score enters the household and replaces its record combination that was checked at an earlier stage. This results in a new combination of the household, which is then checked against the household edit constraints. By sequentially replacing record combinations of individuals with the next highest positive final total field score, all of the household combinations are checked until one is found that passes the edit constraints.

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

- 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.

Inconsistent records and records with missing items need to undergo an imputation procedure. Since there are overall few individual and household records that will fail edit constraints because of the large-scale cold deck imputation integrated into the method for constructing the IAF, we can use standard procedures. One approach particularly useful in the Census framework where there are many potential donors, is the NIM methodology (CANCEIS) successfully carried out on Canadian Censuses (Bankier, 1999). In this method, households and individuals failing edit constraints undergo a hot-deck imputation procedure. Potential donor households that have passed all edits are chosen for a failed household having the same characteristics with respect to the matching variables such as broad age and sex distribution of the household and other geographic and demographic variables. For each of the potential donors, fields are identified that differ from the failed household. A donor is chosen randomly for the imputation from among the possible donors that ensure the minimum change principle of the Fellegi-Holt paradigm.

III. IMPROVING THE EDIT AND IMPUTATION PHASE OF SURVEY DATA

13. As mentioned, one of the most complete administrative sources available for use at the Israel CBS is the NPR. Extensive use of the NPR 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 present an algorithm for improving the edit and imputation phase of survey data using as an example the Coverage Survey of the Integrated Census. We use the NPR for error localization, checking inconsistencies and filling in illogical or missing values of 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 item non-response on survey target variables by using more sophisticated modelling based on the more accurate and complete demographic variables and the geographic variables obtained from the survey data.

14. The procedure that was described in Section II will again be implemented for choosing the best value for each variable in the record which ensures a priori that records pass all edit constraints. In addition to the administrative data, there is now the survey data which is treated as another data source in the algorithm. The adjusted weight, however, for the survey data may be higher (or lower) than that of the administrative files to reflect higher (or lower) credibility. The highest scoring records that pass both the individual and household edit constraints enter the final file of the survey. Only households and individuals that do not pass the edit constraints at the different stages of the process will undergo imputation, thus the number of records that need to undergo stochastic imputation is minimized.

15. The Coverage Survey for the Integrated Census will cover about 10% - 20% of the IAF. The example used in this paper to describe the method is based on the pilot of the Coverage Survey which was carried out in one town in Israel with about 50,000 inhabitants. In this example, the town was mapped into enumeration areas with about 50 dwellings in each area. A 20% area sample was selected which 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,

as well as probability matching on the residuals from the exact matching based on first and last names. The remainder were not linked to the NPR because hey 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.

16. 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 for 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.

17. 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 is included in the edit rule, and zero if the logic proposition is not included 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 is then imported into a SAS file.

18. 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):

- A SAS program transforms the logic propositions into Boolean logic statements (if-then-else) and defines a new SAS program which is then applied to the records in the dataset.

- As a result of running the new 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 or 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 $\times$ 34 propositions) is multiplied by the edit rules matrix (34 propositions $\times$ 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 from 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 Coverage Survey data before incorporating the NPR administrative data and the results are presented in Table 5.

Table 5: Failed Edit Rules for the Coverage 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.

19. By incorporating the NPR data we can correct a priori failed edit rules by choosing the best values of the variables that ensure 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 example where there are only two data sources (survey data and administrative data), a discrepancy in one variable on the record will cause two records to be constructed, each one having a 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. 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 6.

Number of Fields 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

Table 6: Number of Records and Record Combinations for Fields with Discrepancies

20. 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 field score is calculated as described in paragraph 8 which represents the validity and the reliability of the record combination. The total field score is the sum of single field scores which are calculated for each one of the nine variables that differ on the record combination. In general, the single field scores depend on weights that are defined by the user as described in paragraph 8.1 according to the variable and the source of the data in the record combination. In this small example, it was decided that more weight would be given to the demographic variables on the NPR as compared to the survey data. In addition, all variables on each of the data sets have the same weight. Thus, all values of variables coming from the NPR received a weight of 0.6 and all values of variables coming from the survey data received a weight of 0.4. The single field score for each of the fields where there are only two sources of data and all fields in each source have the same weight is trivial and is equal to the weight itself. Fields with no discrepancies in the values between the NPR and the survey data have a single field score of 1.

21. The record combinations underwent the full edit checks. After selecting the records with the lowest number of edit failures and the highest total field score the results in Table 7 were obtained.

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
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%

Table 7: Failed Edit Rules for Record Combinations and Records with Highest Total Field Score

For the records that had the highest total 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 18. Recall that edits E_3, E_6, E_{12} and E_{14} were dropped in paragraph 18. Based on the edits in common, only 44 records had at least one edit failure after incorporating the administrative data using the above method 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 (including edits E_3, E_6, E_{12} and E_{14}), 8,165 records had no edit failures, 1,255 records had one edit failure and 2 records had two edit failures. Table 8 presents the source of the data that was selected for building the final records of the Coverage Survey. Note that for each one of the variables the NPR data had a larger weight than the survey data and therefore the NPR variables were mostly selected for producing for the final record.

Table 8: Source of Data Chosen for Final Records of the Survey

Field	No Discrepancy	Discrepancy between Survey and NPR File		
		Total	NPR Data Chosen	Survey Data Chosen
Sex	9,232	190	190	0
Date of Birth	8,905	517	516	1

Date of Immigration	2,467	6,955	6,924	31
Country of Birth	7,713	1,709	1,615	94
Marital Status	9,019	403	308	95

22. 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 matching demographic and geographic variables. In this example, the few records that had inconsistent data were corrected manually. For the large scale Coverage Surveys, an approach such as the NIM methodology (CANCEIS) mentioned in paragraph 12 will be investigated.

VI. CONCLUDING REMARKS

23. Because of the large cold-deck module that ensures that individuals and households in linked administrative data and survey data pass a priori edit constraints, the scope of actual hot-deck imputations needed at the final stages of processing is limited. This is the main advantage when incorporating administrative data into the edit and imputation processes as compared to conventional methods on survey or census data. We have shown in this paper a methodology for using 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 data processing. Incorporating good sources of administrative data into the survey processing improves the error localization problem, the imputation models that can be used and the quality of the statistical outputs of the Agency.

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.

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 .

Shlomo, N. (2003), "The Use of Administrative Data in the Edit and Imputation Process", UN/ECE Work Session on Statistical Data Editing, Madrid, Spain, October 2003, www.unece.org/stats/documents/2003/10/sde/wp.30.e.pdf

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.

Yitzkov, T. and Azaria, H. (2003), "Record Linkage in an Integrated Census", FCSM 2003 Research Conference, Washington DC, November 2003, http://www.fcsm.gov/03papers/Yitzkov_AzariaFinal.pdf.

Yitzkov, T. and Kirshai-Bibi, N. (2003), “Demographic Characteristics of Non-respondents to the Family Expenditure Survey”, Internal Report Israel Central Bureau of Statistics, May 2003.
