

STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 16  
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing  
(Prague, Czech Republic, 14-17 October 1997)

Item 4 of the provisional agenda

## **EVALUATION OF DATA-EDITING USING ADMINISTRATIVE RECORDS**

Submitted by the Central Bureau of Statistics, Israel<sup>1</sup>

---

<sup>1</sup> Prepared by Olivia Blum.

## **A INTRODUCTION**

1. Evaluation of data editing by using administrative records is a normal routine in evaluating population coverage and response quality in censuses. However, when the editing of a census file, or any other data-file, is already supported by administrative records, these records can no longer serve as an independent external reference for evaluation purposes. Moreover, the required evaluation is not only of the editing process and results, but also of the quality of the administrative register itself.
2. In this paper I would like to address three questions:  
*How is an administrative register used for editing purposes?*  
*What are the advantages and disadvantages of supported editing?*  
*What should be evaluated subsequently ?*

## **B. EDITING WITH THE SUPPORT OF AN ADMINISTRATIVE REGISTER**

3. Administrative registers are data-files that were produced to serve the different interests of private and public institutions. Therefore, their use by statistical agencies for editing data-files, is not a straight forward operation, regardless of whether it is done on a micro or a macro level, within an individual record or of an aggregate. As a matter of fact, the distinction between micro and macro editing, when using administrative register is not as clear as expected. Editing is always carried out on a macro-level, in a sense that it is guided by general rules and conducted on a set of records rather than in a separate and singular process within each record. However the end result is unique to each record because of its dependency on the essence of the corresponding administrative record that is used in the editing process, and therefore it conveys characteristics of a micro-editing operation.
4. Editing with the support of a register usually implies adding or replacing values in the edited file by the parallel value in the corresponding administrative record. It is used in diversified editing tasks:
  - clarifying the identity of an individual record
  - adding register's values to empty fields within record
  - solving problems of out-of-range values within fields
  - solving logical contradictions between fields
  - improving the selection of "sort variables" (of potential donors) for imputation
  - adding new variables to the data-file.
5. Clarifying the identity of an individual record is a by-product of the record linkage operation. Once the records of the same individual in both files are linked, the editor can add or correct identifying variables in the edited file. For example, when the individual's unique ID number in the survey is not identical to the one in the register (short of control digit etc.), but a linkage is possible based on geographic and demographic variables, the editor can correct the ID number in the survey data-file. By doing so he confirms the mere existence of the individual whose record is linked and the record's identification.
6. Empty fields are at times left as missing variables, but when a statistical editing process is taking place they are replaced by values reached by different methods (central tendency, linear/nonlinear model, imputation etc.). When an administrative register supports the editing process, empty fields can be replaced by the value found in the register. This process is beneficial if and only if it does not create a new editing problem in the edited record.

7. Out-of-range values within fields are usually edited by erasing the value and treating it as a missing value.

8. Logical contradictions between fields require a more careful treatment. In statistical editing, when more than one field is involved in an editing problem, editors have to find out which one is not supported by other variables in the record and focus on it while editing. When it is not possible to reach a decision, editors either erase all values or leave them as is. An administrative file can serve in such cases as an external reference that can help not to refute but rather to corroborate the values in one or more fields involved in the failed edit check. Editors can refrain from editing a field that has identical value in the register, and replace a field value by the register's when values in the mutual field are different. Here again, the operation is worthwhile if the replacement settles the logical contradiction.

9. In the imputation step, the selection of "sort variables" of the potential donors is usually confined to the relevant variables within the edited file. If an administrative register holds additional variables that can refine the selection and enrich the pool of possible donors, it is worthwhile to consider doing it. Costs of an overall record linkage that is needed for such operation, should not be ignored.

10. Although it is not an editing task per se, adding new variables to a data-file is frequently done in the editing step. Some of the new variables come to serve better management and control of the file (statuses, dates, consecutive record numbers etc.). There are also new calculated variables (age that is based on date of birth, number of accumulated years of schooling, affiliation to the labor force etc.). Record linkage with an external file broadens the possibilities. For example, in Israel's 1995 census of population, no questions about religion were asked, however, in the completed census file each individual record has this variable, that was taken from the population national register.

11. There is another use of a register in an editing process, that is not done on an individual level, but rather on aggregates. This process is based on statistical record linkage (of groups) and on comparisons of distributions. I do not elaborate this point in this paper but it is a worthy issue, especially in cases where individual record linkage is not possible.

### **C. NECESSARY CONDITIONS**

12. Not any administrative file can be used for editing purposes. Since the main feature of editing with an external support is replacing erroneous or missing values in one file with values in the other, the ability to link individual records becomes crucial to the process. This characteristic introduces several conditions to the use of registers for data editing:

13. The file to be edited and the register have to have common variables, otherwise no linkage will be possible.

14. Both files have to have a set of identical identifying-variables for record linkage that serves for micro-editing purposes (within records). A unique ID number is not always enough for record linkage, because of the possibility of erring while collecting data, therefore the linkage has to be based on an elaborate identifying profile.

15. The number of common variables have to exceed the number of variables needed for record linkage, or else, the register would have no added value and therefore should not be used for editing.

16. The common variables should be defined similarly. These variables, coming from different sources, are usually collected in a different manner, by different agencies and although their titles are the same, it is not clear that their definitions are identical. There is a need to find out what a value in a certain field represents and whether it is possible to compare it with the respective value in the administrative file.

17. The number of mutual records have to justify the complex editing operation. There are designated files that have records of only part of the population like health insurance, vehicle register etc. Even if these files are of high quality and interest, if they hold records of only a small portion of the population in the edited file, their use would be too expensive and their contribution too small.

18. The structure of the administrative register has to be "editor-friendly". If there is a need for human and machine resources in order to prepare the register for a secondary use, one should consider cost-benefit ratio of such an operation (is it done for a single or multiple use, are there any alternatives etc.).

#### **D. ADVANTAGES AND DISADVANTAGES**

19. In order to examine the desirability of the use of administrative records for editing needs, one should consider comparing it with statistical editing. The advantages and disadvantages of using administrative records are as follows:

##### **Advantages**

20. The use of administrative files enables the editor to identify patterns of erroneous responses in the census data file, regardless of their origin (whether in the data collection step or in a systematic response bias), even if consistency errors are not detected. In a way, this is a process of embedding an evaluation of the file, while editing it. Census files are evaluated by comparison to demographic estimates while here, editing is done through comparison.

21. Register-supported editing also enables imputation by register values that even if they are not representing an objective truth, are, at times, preferable to missing values or statistical imputation. The register may present a different picture, but it is more of an empirical nature than statistical editing.

22. The use of administrative records means conducting macro-editing that is advocated by micro-editing. Each record enjoys individual editing by getting the values given for the same individual, in the same field, but for other purposes. Yet, the editing process is of a macro level nature.

##### **Disadvantages**

23. When qualifications for benefits are stipulated by the data in the register, this file is biased toward the qualifying attributes. For example, if taxes are lower in selective parts of the country, people who move to a higher tax zone would not change their address in the register, unless they have some other incentive to do so. Although editing problems are solved by the register values, the edited picture may carry reliability problems caused by socio-economic interests.

24. Not all registers are updated at real time. When the content of a register is a function of individual's reports in addition to administrative practice, cultural variables as well as interests are involved. Cultural patterns mitigate how reliable the population reporting pattern will be. For example, there is a difference in reporting patterns between people in societies that lean heavily on improvisation in their everyday life, and those who are used to educated planning, between people in conservative societies and those in societies of liberal milieu, etc.

25. The editing process that uses administrative records is based on record linkage and therefore introduces the possibility of wrong matches. This occurs when a record has only a partial profile. This phenomenon is observed in homogeneous groups such as foreign residents who do not have or do not supply the information needed, and also when there are ideologically motivated refusals to answer the questionnaire.

26. Finally, the editing process does not make an overall use of the registers but rather a selective use in cases of failed edit checks. The population involved in actual editing has to be analyzed and characterized beforehand, in order to free the data from a possible bias.

## **E. EVALUATION FEATURES**

27. The target file for evaluation is the edited file, i.e. the outgoing product of a census or survey. During the evaluation of the editing process, it is broken down to sub-processes in order to identify source, in addition to direction and intensity of a found bias. Since the administrative register adds its own biases and errors, evaluation cannot be restricted to the edited file but rather include the register itself.

In this section I refer to evaluation features that are introduced by the use of administrative register. I do not elaborate on evaluation of editing per-se.

28. The basic operation that enables the evaluation is a comparison between all records of the edited file and the register. It produces different data groups that serve the evaluation purposes: edited data and control groups of unedited data, linked and unlinked records, census or survey surplus records and register surplus records.

29. In order to be efficient in evaluating editing, a pre-screening of the edited variables is needed. Each variable should be characterized by the intensity of the use of administrative register for its editing. The decision of which variables to include in the evaluation process should be made accordingly.

30. There is also a need to characterize the population involved in the edited records, in order to find out if it is significantly different from the population in the unedited records. A significant difference means that the edited file should be evaluated not only directly but also indirectly by the evaluation of the register. It may lead to an evaluation by a third file in addition to internal comparisons.

31. A third data-file (of surveys or demographic analysis) should be also used in cases where the proportion of the linked records is not very high. Large proportion of unmatched records decrease the ability to evaluate the mere compatibility of both files. They may represent different populations altogether.

32. Although population coverage and response quality are the main evaluation components, the involvement of the administrative register in the editing stage introduces

additional editing segments to evaluate: record linkage, deletion of fictitious duplicate records (a result of wrong record linkage), surpluses of both files (unmatched records), variables' definitions, content error in the register (and therefore in the edited records).

33. All editing and evaluation procedures are based on record linkage, hence, it is the first evaluation objective. It can be done automatically by activating different criteria for linkage, except for the one used for actual linkage, or manually by sampling cases and checking the identification of the individuals involved.

34. Evaluation of deletion of duplicates should be done together with the previous one, using the same process of testing the linkage with different criteria. When the record linkage is wrong, two records have the same ID number although they belong to two different people. When a deletion of duplicate records is based on the ID number, it leads to deleting true records and consequently to under-coverage caused by editing.

35. Coverage evaluation is done on a local geographic basis. In the boundaries defined for coverage evaluation, linked files of census and administrative register have three types of records: linked records, register surplus and census surplus. The unmatched records are the main input for coverage evaluation. Some of these records are not matched because of technical problems, such as missing data, but actually represent records of the same individuals. Some of the unmatched records are a result of erroneous records-deletion in the editing stage. The other unmatched records indicate under or over coverage of the files involved.

36. There is a problem of deciding what file indicates the truth; does a surplus in one file is an over-coverage of the file or an under-coverage of the other. A third file can be useful if its variables supplement missing information. For example, when an individual record exists in the register and does not exist in the census, a "deaths register" may indicate that since the person died, the administrative register is not updated. In that case, the census does not suffer from under-coverage. When the third file does not hold information of that kind, it can only increase the probability that one of the files is more right than the other. If no data are decisive enough to conclude one way or the other, and the problem is of a large scope, a designated coverage-survey should be initiated.

37. Differences in terminology and in variables' definitions between the register and the edited file influence the response variance and contribute to editing dependent biases. It should be evaluated by indices of inconsistency.

38. Content error in the administrative register has to be estimated by comparing it to "true data". A third file of a designated survey or of a survey conducted carefully for other purposes, can serve this objective.

However, a comparison between the register and the edited file can be part of the evaluation in itself; Comparing populations in unedited records with the corresponding ones in the register, means that inconsistency indices can be measured both ways; In one process the unedited records serve as the true file (in selective variables) and in another the register does. Significant difference between the edited records and the unedited ones in the survey should be taken into account.

## **F. CONCLUDING REMARKS**

39. The empirical work of the evaluation will be the topic of another paper. However, the theoretical considerations described above are relevant for more than mere empirical planning. The physical and logical handling of two files that hold records of the same people and their evaluation, is applicable to the mechanism of record linkage of administrative files, and to their use as an alternative source of information to conventional surveys and censuses:

40. In order to use registers for editing other data-files (surveys and censuses) of other registers (for administrative census and such), there is a need for standardization of concepts and definitions, intranationally and internationally.

41. For the benefit of editing aggregates, techniques of aggregative record linkage should be practiced. When individual record linkage is not possible, but there are indications that there are groups of unidentified or unlinked individual records, in both files, that belong to the same population, editing should be of aggregates and should rely on attributing characteristics to groups rather than individual records.

42. The problem of deciding what source of information is more reliable is yet to be encountered. Registers evaluation is the most prominent and critical operation in a transformation phase from conventional to administrative censuses. However, there is also a need for routine evaluation along the process of administrative data collection in order to define which file to use as a core file and which file to use for editing purposes.