WP.16

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 (ii): Implementing editing strategies and links to other parts of processing

GLOBAL AND LOCAL OPTIMIZATION OF EDITING: A PRODUCT-ORIENTED APPROACH

Supporting Paper

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

I. INTRODUCTION

1. Planning editing and imputation is complex and challenging when data are of different sources and are collected in an inter-dependent process. Each source of data carries its own unique errors, the mere integration of data is a source of errors in itself, and when the integrated file is an intermediate product, more errors are accumulated before the final file is sealed. Moreover, errors in one phase of the production process stipulate the scope and content of editing needed in the following processes. Hence, errors have to be treated during the data collection process, during the integration of files and after each manipulation on the data. In this course of action, the active ensuring of the appropriateness of the data is done simultaneously with the data-source selection, i.e. editing and imputation serve not only as a validation and correction phase, but also as an integral part of the data collection process itself.

2. In this intricate and tangled process, where data are used for different purposes, an overall editing strategy is required. The inclusion of an evaluation that leads to errors identification and localization, and the inclusion of quality assurance that control or correct errors along the statistical production process, implies a product-oriented approach. The required quality of the product is set as the goal, toward which editing processes are planned.

3. Since resources are limited, the overall editing plan should be optimized. A product-oriented approach entails a global outlook, where optimization defines the most efficient way to the required final-product. However, in a multifaceted project, where the statistical quality of the intermediate products has a critical impact on the quality of the final product, optimization of editing should be local as well as global. Quality thresholds are to be met by the intermediate products in order to be able to have a qualitative final product. For example, an institutions-register is a prerequisite for the enumeration of residents of communal living-quarters. In a nontraditional census, when only part of the area is covered physically, coverage of all institutions stipulates the coverage of all residents in their census-addresses; therefore, the completeness of the register, the intermediate product, is set as a goal to be reached in the editing plan.

4. Intermediate products present additional challenges to the editing plan on the global as well as on the local level. These products are used, at times, for more than one final product. The institution register, for example, is also used for the current small-area demographic-estimates. This use requires a broader editing plan that optimizes additional quality dimensions, like accuracy of age and gender. The implication is that the boundaries of the global editing plan of a specific project have to be well defined beforehand, and that it may include processes not directly related to the project at hand. On the local

[1] Prepared by Olivia Blum blum@cbs.gov.il

level, the intermediate products are, at times, final products for themselves. The institutions-register in the example is a source of institutions statistics, needed for planning and implementation of social policy. As such, it requires more extensive editing to serve the quality needs of the users. Hence, the definition of internal boundaries is essential as well.

5. The next Israeli census of population is an integrated census that uses several sources of information and serves several intermediate and final products. In the following sections, the editing requirements and the suggested strategy for implementation are presented.

II. PRODUCTS AND EDITING CONSIDERATIONS IN THE ISRAELI CENSUS

6. The next Israeli census of population intertwines editing processes throughout the statistical production process, from the early stage of data-source selection, through data-collection and integration, up to the formation of the final census files. Administrative records are combined to form a geo-demographic file, and sample surveys are conducted in order to supply small area estimates for individuals and households, characterized demographically and socio-economically. Moreover, several intermediate products stipulate the quality of the final census files, and are used for other final products or as final products for themselves.

7. The following table presents the main intermediate products and the related processes that involve editing:

Intermediate Products	**Processes**
Evaluated administrative data-files	· Evaluation of · File · Variables · Records · Selecting the relevant files.
Edited administrative data-files	· Deletion of failed items and units. · Geospatial-coding. · Harmonization.
Integrated administrative file (Edited while integrated)	· Record linkage. · Building the frame and defining the analysis units. · Selecting the residential addresses. · Selecting values of other variables.
Field-surveys files (Micro-edited while captured)	· Computer-assisted data-collection. · Management of data-collection.
Edited survey files: Geo-demographic, Socio-economic	· Macro-editing. · Coding (addresses, occupation, industry).

8. The accessibility to administrative files that are suitable for census purposes is limited and usually involves costs. Therefore, a source-selection process is needed and it is based on a careful evaluation of the statistical quality of the contents and the physical usability. The data-file is evaluated in terms of the population and the variables that it covers, and the coherence of its contents. Specific variables and records are also evaluated; variables definitions and categories are compared to the required census definitions, and the completeness and the coherence of special target-population records are measured. This evaluation is functional to the census needs, and even if the administrative file is problematic from different aspects, it still may be usable for the census. For example, the electric company consumers' file carries, at times, the details of the owner rather than of the person who lives in the address, with no indication for that. One person can have more than one address in the file, in different records that represent different electric consumption units and different dwelling units. In some cases, it is not possible to geo-code the address, or to link records with the corresponding ones in the population register. However, this file has been found to be better than other available administrative files to replace the addresses of the population register, that are set as a default, in cases of disagreement between the two. In an experiment conducted in a medium-size town in Israel, 12 out of 16% of disagreements have been found to be correct addresses in the electric company file (Among the 84% of agreements, 3% are erroneous).

9. In addition to being a potential source of census data, the administrative data are also evaluated for other statistical purposes in the office (See also: Blum, 2005). For example, updating the current demographic estimates is based on the accumulation of the changes, registered in a defined time-unit, in the satellite files of the Central Population Register (CPR). The demographic estimates are not a reflection of the CPR since it carries historical errors. The starting point of the estimates is the last census, and the changes over time are those of the CPR. Lately, the demographers have contemplated the idea of extracting the changes from the CPR itself and not from the files that record changes only (births-file, deaths-file, changes of marital status, etc.). The added use of the administrative files requires additional evaluation and hence, additional editing.

10. Evaluation of data files starts a long time before the census and it is repeated in the experiments intervals. An updated version of the selected files is used for the experiments and eventually for the census. As a preparation for the incorporation of the selected data into the integrated census file, the separate files are functionally edited. Editing tasks include harmonization of the relevant variables along the census definitions, detection of errors, deletion of failed units and items, and geospatial coding of addresses. Geospatial coding is used for several processes of great importance in this census, among which is linking records of different spatial entities and area sampling.

11. The next census in Israel is administrative in its geo-demographic part only, and it relies on the CPR that serves as a default file. The integration of the other data files serves mainly the definition of the frame and the analysis units, globally and locally, since addresses of people who move within or out of the country are not updates. All files are linked on the record level, using shared PIN, coded addresses and demographic variables. Records of suspected emigrants are tagged as such, based on a probabilistic model. Local addresses are determined through a selection process, which is basically a cold deck imputation, where the most qualitative source of data contributes the address or its geo-spatial code.

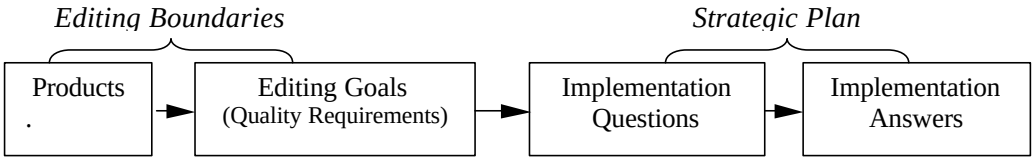
12. Individual and household are the basic analysis units of the census. The records in the integrated file represent the individuals, and by the end of the census, when weights are assigned, each record may represent a fraction or more than one person. The definition of households is more problematic since addresses are not detailed to the apartment level, but rather to the building level. People, who have family connections to other people living in the same building, may form one or more households. Moreover, people with no family connections may do form a household. Therefore, household statistics is drawn from the field survey only. Yet, a family is defined as a proxy of a household in the integrated administrative file, since it facilitates fieldwork operations in the census surveys and supports geo-coding of addresses. An address of one family member is edited with the support of another's.

13. The sample surveys provide the census with the socio-economic information and the final demographic estimates. Census surveys are too complex and costly to serve other statistical projects in addition to the census itself. Functional editing is restricted to the census only. Micro editing is carried out during the computer assisted data collection, within the questionnaire and outside of it in the management supporting processes. After the data collection, macro editing is performed. It is functional since it serves a well-defined purpose. For example, if the only purpose of an operation is to find out if a person lives in or out of the statistical area that the estimates relate to, the number of the building becomes irrelevant if the whole street is within the area. In such streets, numbers are not edited. Moreover, if variables are collected only to enable record linkage, once the records are linked, those variables in the survey file become irrelevant and are not further edited.

14. The final census products are the geo-demographic file and the socio-economic file. The first is a weighted integrated administrative file, calculated by the surveys estimates (See also Nirel, Glickman and Ben-Hur, 2004). This file is mainly used to supply the demographic margins of the census. The second file is the one that the vast majority of the users use. It has all census information of about 20% of the population. It is the source of socio-economic and household data, and its demographic information, although not exclusive, is required in conjunction with the other attributes. This file is adjusted to the demographic benchmarks, provided by the geo-demographic file, by weights and imputations, and has to meet the quality requirements of the census.

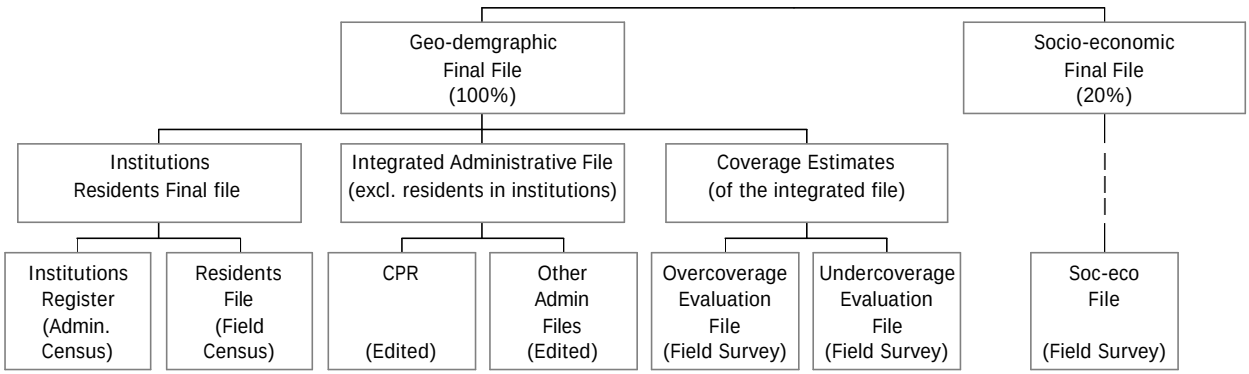
III. EDITING OPTIMIZATION BY PRODUCT

15. Planning functional editing in a product-oriented approach includes the definition of all products, within the main project and those related to it; their required quality; and the derived questions of implementation. The products and their respective quality requirements set up the boundaries of the editing task at hand, while the answers to the implementation questions are translated to an overall process and are expected to lead to an efficient and parsimonious plan.



16. Since products of all levels are introduced, editing processes of hierarchical order are involved, i.e. no contradicting processes are expected, and the global optimization of editing is relatively straightforward. However, when the intermediate products are final products for themselves, or they lead to more than one product, internal or external to the project, optimization of editing is of simultaneous processes, that may have requirements of different nature, and therefore, local optimization of editing is necessary (Summarizing table in annex 1).

17. Following the above model, the following diagram is of the census main intermediate-products and most connections between them:



18. The editing plan toward each of the two final products is utterly different: Editing data collected to form the socio-economic final file is an example of a global process with no need for local considerations. The data collected in the field survey is the only source for socio-economic characteristics and for household composition. There are no intermediate products or external products to edit toward. The only external imposition is the weights that have to be added to the records in the final file, because of the sample and according to the demographic benchmarks, derived from the geo-demographic final file. A need for strategic editing plan is demonstrated in the processes of building the geo-demographic final file. It is a three level data: Basic files are integrated to form the intermediate products, which are combined to form the final product. Although the production process is incremental, the planning of editing in a product-oriented approach goes in a reverse order from the final product to the basic files, as demonstrated in the following example:

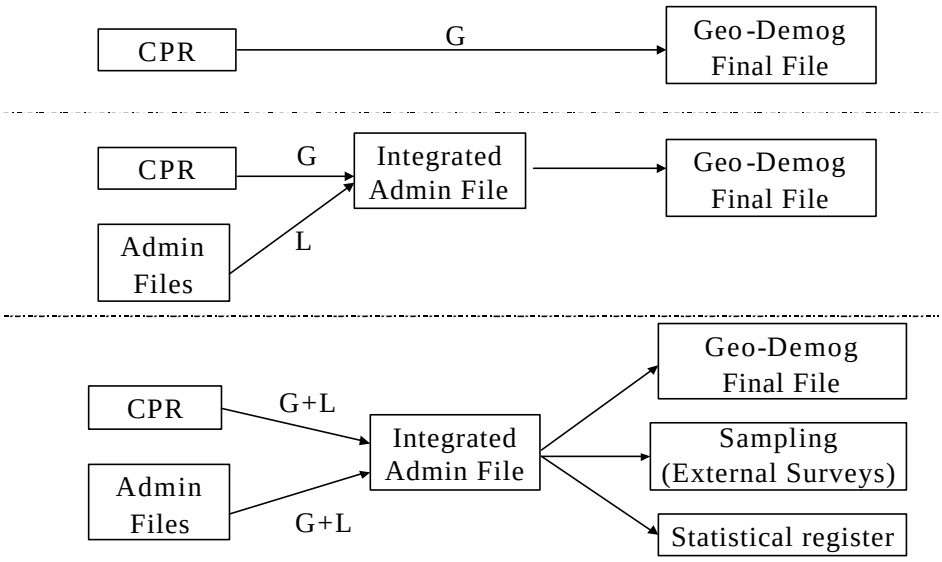
19. Defining the frame of the population is a major census challenge and it goes back to the fundamental census goal of counting the people. The default frame of the integrated census is the Central Population Register. It could have been the census frame if quality requirements have been met. However, about 10% of the records are of people who reside abroad or even died there, while illegal foreigners are not covered. Focusing on the emigrants, four types of information are used to identify them and estimate their size: Prediction of the probability to be an emigrant, based on an 'emigrants model' that uses data from administrative sources, an institutions' residents file based on a full census, an

undercoverage survey that find a sample of the wrongly suspected records, and overcoverage survey that looks for the people not found in their CPR addresses.

20. From editing perspective, global optimization of the process in this example leads to two considerations only: Enabling record linkage and an imputation of the status of belonging to the census population. There is no need to edit the basic files for themselves. Moreover, the overcoverage survey has an embedded link to the CPR since it is the source for the identifying variables of the sampled unit, hence, the only information needed from this file is whether the person lives in the country or not.

21. However, the very same basic files are used for other purposes as well. The advantage of a census over a sample survey is the ability to supply small area estimates. It turns to be problematic since about 30% of the addresses in the CPR are not updated. The extensive use of other administrative files for cold-deck imputation of addresses is the chosen solution, and the resulted product is not an edited CPR but rather an integrated administrative file (See also: Shlomo, 2002). It is not the final geo-demographic file since it lakes the weights acquired in the coverage evaluation surveys, and the institutions' residents that are fully enumerated. However, the inclusion of a new intermediate product brings about new uses and therefore a need to plan editing accordingly. The integrated file is a relatively good source of addresses of sampled units in external surveys. It is also a reference file for an external demographic project that tries to maintain a statistical register of the population on an ongoing basis. Editing toward these uses is locally optimised. The external surveys may have to use additional addresses in the different administrative files, therefore, these addresses have to be edited not only for address selection in the cold-deck imputation, but also for unique identification of the geographic location, since an interviewer has to reach the dwelling.

22. The following diagram illustrates the changes of the editing orientation, with the introduction of new products, intermediate or external ones:



G-Global editing; L-Local editing

23. Once the boundaries of the editing task are defined, implementation questions arise. These questions refer to the overall plan, to the data-files to be used, and to the variables. Continuing the addresses example, a question of an overall plan is whether there is a need to define intermediate products like the integrated administrative file? Should the products serve other purposes? What purposes are legitimate to include and what are to be kept out of the editing scope? On the data-files level, questions that guide the evaluation process are to be asked: Is the file useful (Improves more than damages) and should be a used as a source for address information? What is the preferable order of the files to be used for cold-deck imputation (What is the marginal benefit)? Is there a need for an intermediating third file to enable record linkage between two files? Should the ability to geo-code addresses be used as a criterion for the file-quality? As for the variables, the questions are closely related to the selective editing to be implemented: Should geospatial entities used for record linkage? What supporting variables, which are not included in

the final files, should be edited and how? The answers to the questions above are pre-requisites for a functional editing plan. They stipulate the type of editing to be engaged in the different parts of the production process.

IV. CONCLUDING REMARKS

24. The statistical production becomes more and more complex since it uses many sources of data and it leads to many end products. Consequently, editing has to widen its scope to include these ends and to encompass data manipulations that have not been part of editing in the past. Furthermore, a product-oriented editing-strategy is called for, because of the quantity and the diversified nature of the processes involved. In this realm, optimization of editing cannot be done on a global level only, but also on a local level, where the quality of intermediate products is set as a goal as well.

25. Extended scope and diversity of editing processes may eventually lead to a complicated or even inapplicable implementation, under given resources. There is a need to identify the threshold beyond which excessive editing, carried out under simple universal rules, becomes less damaging than the selective, product-oriented, specific editing. This threshold should be defined, inter alia, in terms of the number of data files, variables and meaningful products involved, the marginal costs added in the transition to undistinguishing editing process, and the applicability of the plan. Beyond the threshold, a pure product-oriented strategy will not prevail and a different approach is expected.

References

Blum, O. (2005) “Evaluation of Editing and Imputation Supported by Administrative Records”. UNECE. Work Session on Statistical Data Editing. Ottawa, Canada.

Nirel, R., H. Glickman and D. Ben Hur (2004) "A Strategy for a System of Coverage Samples for an Integrated Census." Proceedings of Statistics Canada Symposium 2003: Challenges in Survey Taking for the Next Decade.

Shlomo, N. (2002) "Smart Editing of Administrative Categorical Data”. UNECE. Work Session on Statistical Data Editing. WP21. Helsinki, Finland.

Annex 1

Table of variables within global and local editing goals in the integrated census (partial list):

No	Intermediate Product	Aimed-toward Product	Quality Goals Completeness and Accuracy of -
1	Institution register	Geo-demographic final file	Global: Addresses and number of residents
		Residents' file	Local: Identified records, addresses and types of institution.
		Itself	Local: All records and variables.
		Current demog. estimates	Local: Addresses and number of residents
2	Residents' census file	Geo-demographic final file	Global and local: All records and variables
		Inst. residents' file	
3	*Inst. Residents final file*	Geo-demog. final file	Global: Frame (all records of institutions' residents) and location.
4	CPR	Geo-demog. file	Global and local: Frame (all records of census population), all identifying and demographic variables, addresses.
		Integrated admin. file	
		Current demog. estimates	
5	Other admin files, like the electric and telephone companies, previous census	Geo-demog. final file	Global: Identifying and record linkage variables, addresses
		Integrated admin. file	Local: Identifying and record linkage variables, addresses, families
		CATI surveys	Local: Telephone numbers, families
6	*Integrated admin. file*	Geo-demog. final file	Global and local: All variables
		Demographic project (statistical register)	
		Sampling for external surveys	Local: Identifying and record linkage variables, addresses.
7	Overcoverage survey file	Geo-demog. final file	Global and local: Addresses
		Coverage estimates	
8	Undercoverage survey file	Geo-demog. final file	Global: All variables
		Coverage estimates	Local: Identifying and record linkage variables, addresses
9	*Coverage estimates*	Geo-demog. final file	Global: Addresses in each statistical area
10	*Geo-demographic final file*	*Itself*	Global: Weights
11	Socio-economic survey file	Socio-economic final file	Global: All variables
12	*Socio-economic final file*	*Itself*	Global: Adjustment to the demographic margins (weights and imputations)
