WP.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (iv): Editing of Census and social data

AUTOMATIC DATA EDITING EXPERIENCE IN 2010 MEXICAN CENSUS

Prepared by Isaac Salcedo, INEGI, Mexico

I. INTRODUCTION

1. The Population and Housing Censuses are the largest statistical project in any country and provide highly relevant information for: policymaking, research and planning, so its results should be disseminated with the greatest opportunity which is one of the reasons why the data editing must be a faster process inside the all global project.

2. This paper describes the automatic data editing experience in the Mexican Census whose efficiency impact the lead timely results. The methodology is described with particular emphasis on review editing results strategy, carried out at the local offices of the Institute, whose maximum level of disaggregation was the municipality.

II. DATA EDITING

3. The objective of the data editing is improve the information's consistencies, because the data coming from the work field, information entry or coding have some problems in its congruence. That is, assure the quality of information for the generation of statistics according to a conceptual framework. And as result of this, the data editing provides important information about the data quality and improvement of the questionnaire.

4. Derived from questionnaire there is a logic relation between questions and sections, however, as mentioned above during field processes and data processing mistakes and inconsistencies could be generated. Four kind of errors must be treated:

- **Missings values**: It is the absence of information in a question, where according with conceptual design must be exists.
- **Multiresponse**: It occurs when a question has more responses than permitted in its design.
- **Inconsistent values**: When exist a response in the question but is not a valid value for that item or is not logical with other answer related with it.
- **Ignore filter questions**: This mistake occurs when some responses should not be registered in a question because there exist a filter.

III. THE PROCESS STRATEGY

5. The data editing strategy included three major stages: design, training and implementation, the first one was developed form January to July, the training for staff was during the last two weeks of July and finally the implementation of census data editing was conducted between August and December 2010. On December 6, 2010, census information was released from the automatic editing phase.

Design
Training
Implementation

Figure 1. Major stages in data editing process.

6. There was specialized staff in central and local offices for this process. This staff were hired and trained specifically for this project, in the table below each structure functions are presented.

Central offices	Local offices
-Design of data editing strategy and reports	-Receive training on automatic data editing
-Design and review of data editing criteria	-Train lower level staff
-Programming of criteria and reports	-Review the reports of automatic data editing
-Development of manuals and training strategy	-Report weekly results of the analysis
-Provide training for local staff	-Formalize the completion of the local data editing process
-Supervision of local staff	
-Formalize the completion of the national data editing process	

IV. DESIGN

7. Undoubtedly the design stage is the most important of the whole process, main lines of action were defined in this step as well as the functions that perform the figures involved, the definition and programming of editing criteria, reports that will be analysed (those are essential in the process) and parallel training was prepared, that is manuals for each of the sequential steps to ensure that everyone involved in the process had clear dates and their functions.

8. The automatic data editing was developed development considering this guidelines:

- Keep the conceptual and methodological standards established for each issue and variable.
- Preserve the correctly information collected.
- Apply a review of logical consistency supported by empirical evidence observed in previous projects.
- Assign values as long as they are supported by the information contained in other associated variables or impute the missing information.

A. Definition of editing criteria

9. For the criteria design were analysed the 2000 census criteria, in order to take advantage of the previous experience. They included criteria editing derived from direct processing and others designed

based on the methodology known as theoretical vectors. However, the design of these editing criteria does not guarantee that a variable is not affected by two different criteria. It is noteworthy, that the use of this technique is preferred because it provides some advantages such as:

- The information analysis is exhaustive.
- It's more powerful than the flowcharts technique.
- This technique make easier the analysis and tracking changes.
- Support programming.

10. The theoretical vectors methodology is a tool that allows complete control of the values that can take a set of conceptually related variables, the procedure for its application is as follows:

(1) The variables considered in the editing criteria and the values taken for each one are identified.
(2) A one-dimensional array is constructed (called theoretical vector) whose components represent the values they can take each of the variables.
(3) All combinations that can take a vector are generated, these are obtained by varying one to-one the values of each component of the vector, starting from the last to the first component , to obtain the set of total combinations.
(4) To have total control over the generated combinations, it built a function called routing function, which allows assign a unique value, known as image, for each combination in the generated order.
(5) The result of applying this method is to have all possible combinations between the values of the vectors, in addition to an identifier of each combination and therefore, for every combination is possible to consider a specific treatment.

11. There were three types of questionnaires: basic for complete enumeration, extended for a sample and self-enumeration for extemporaneous enumerations, for each one criteria editing were designed, however for common items the editing was done with the same criteria. In total, 232 criteria were designed for automatic data editing, the distribution is shown in the next table.

Theme	Total of criteria	Theme	Total of criteria
Initial Treatments	7	Dwelling Characteristics	29
Food Access	3	Treatments for population	6
Sex and Age	16	Household Relationship	35
International Migration	20	Entity or Country of birth	4
Health Services	4	Religion	2
Disability	15	Indigenous Language	4
Educational Characteristics	12	Residence Entity or Country in 2005	8
Marital Status	10	Economic Characteristics	30
Fertility	27		

12. Treatments within each theme have an order and were applied sequentially, the order of application of the criteria was important because this ensures proper treatment of census data

B. **Reports**

13. Six types of reports were designed, they allowed to review the editing data was complete and free of inconsistencies. These reports make easer the review of the changes to data during automatic editing and check that those were corresponded to expected changes and the effects on the data were within the expected range for errors or inconsistencies. A brief description of each report and the order in which those should be reviewed is presented below:

(1) **Control of initial treatments:** allows to evaluate the impact of initial treatments, which essentially consist in delete records without information and ensure that the kind of housing is unique.
(2) **Control totals:** presents the total of the main subsets of housing and population, before and after of initial treatment.
(3) **Related variables:** with this report the congruence between two variables is analysed once the editing is complete.
(4) **Origin-Destination matrix:** This report shown changes in the distribution of a variable before and after the editing. It is generated for all variables of all questionnaires and the information is presented in absolute numbers and percentages.
(5) **Missing and not specified values:** This report assesses for each variable the percentages of omission and not specified values for ensure they do not exceed the permitted tolerance.
(6) **Percent changes:** displays a summary of the changes for each particular variable affected for the applying of the treatments.

14. In total there were 379 reports generated at the State level when editing started, but as the data were released of coding process, they were generated at the municipality level. The following table shown the total of reports to be analysed by type.

Type of report	Total
Control of initial treatments	1
Control totals	2
Related variables	247
Origin-Destination matrix	125
Missing and not specified values	2
Percent changes	2

C. Programming

15. Each criterion was programmed separately and was integrated into the sequence previously defined, at this point a critical activity was the test out of the correct programming of all criteria by itself and by the set; for did it, a simulator was developed and was fed with a data base with all possible errors. For a week, it was reviewed in detail the results of the criteria programming, as well as reports generated for further analysis.

16. As mentioned above, local staff had the responsibility of reviewing every report from the data editing, so at the pair of systems development, the central office staff implemented a strategy of continuous communication between the local and central responsibles of the process. For this, formats were designed in which the local chief reported weekly the progress and incidents occurred during the week, for the central staff.

V. TRAINING

17. The training was designed with emphasis to the review and analysis of automatic editing reports. During the design of the editing process, manuals were prepared to facilitate analysis of all reports in order to the staff responsible of its reviewing could correctly interpret each of generated reports. The purpose of each manual was provide the necessary tools for the qualitative and quantitative analysis of the information resulting from the editing process.

18. Training had two big stages. The first one is given directly from designers of the process and was directed to the data edition's responsible, the second one was given by the responsible to the editing analysts.

19. The content of the training was designed in order to the new staff in the Institute had the opportunity to understand the work that it performs like a producer of basic statistical information. The conceptual concepts of each section and question from the three questionnaires were explained. The emphasis was in the explanation on the type of items in the questionnaires and the relationship between each of the sections and questions, even more they filled out questionnaires simulating real situations.

20. About the automatic data editing, the need to perform this step was explained as well as every possible inconsistencies from the data. At this point, many examples were made, in order to have greater clarity and sensitivity about the importance of the data congruence.

21. Concerning to editing criteria and treatments, the theoretical vectors methodology was explained but only a few criteria were analysed with all detail because the principal labour of this personal was the review of each report. However, the complete manual with the editing criteria was provided so it could be consulted if some doubts arose during the review process.

22. For the staff in the local offices, three days of training were destined for the analysis of the editing reports that was in order to explain with the higher level of detail possible the reports. For each type of report, was explained: the content, its integration and how it must be analysed.

23. At the end of the training the people was examined in order to ensure that everybody had the necessary knowledge for their activities and in other hand, the communication strategy was disseminated (it was bases primarily on the use of email and phone) as well as the use of some software that they must use during their work.

VI. IMPLEMENTATION

24. This activity contemplated the automatic editing criteria execution, the generation and analysis of reports to verify that it was properly made and at the end of editing process each State data base started a release process. The complete procedure had a duration of thirteen weeks.

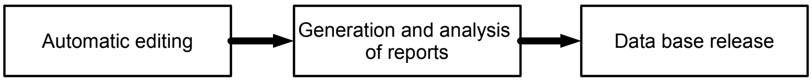

Figure 2. Implementation.

A. Analysis of reports

25. Weekly all editing responsibles from each State sent their reports to central office with the results of the analysis of the automatic editing, the most of their observations were:

- Tolerance for some variables was exceeded, in general this problem was submitted in variables with small population.

- The total of the same population between different reports was not equal, however this observation although was frequent many times was false.

26. Derived from the false observations it was necessary provided backing up on some concepts as the contrast of total population in specific reports and communicate about the correct way to inform. Those actions were very important for all responsables in their performance because quickly they can reduce the false observations in their informs.

27. As a result of the analysis of the reports and the information received by the central office there were adjustments in some criteria of data editing which were studied for the conceptual department before its application.

B. **Data base release**

28. After the criteria adjustment all data were submitted to a complete automatic editing in order to ensure the homogeneity of its applications, this process took 165 hours of processing time. At the end, every state responsible revised the resulting reports as was usual.

29. A second step inside the data base release was the analysis of missing values, this task consisted in a contrast between observed and expected percentage of this characteristics. If some variable exceeded the tolerance was necessary a justification for the variable release.

30. Finally, the editing responsible elaborated a document called "Editing Analysis" in which was detailed the data behaviour before and after the automatic editing, with particular emphasis on missing values and the impact of this process in every variable, in special if a variable had a change higher to one percent it was solicited an explanation of that. Every document, were examined by the staff of central office in order to be sure the consistency of the release database.

VII. **CONCLUSIONS**

31. In general, it could say that the stage of the automatic editing for this census was carried out successfully, presenting a significant improvement compared with previous censuses, mainly in the control of changes in the information and the exhaustiveness of the treatments. This last point derived from the fact that the design of the criteria for automatic editing for this census were developed totality with the theoretical vectors methodology; which enabled have a better control and completeness on the treatment of variables.

32. The time assigned to carry out this project, from design to implementation and delivery of the database released, was significantly lower than previous events. The design stage was decisive for the correct operation of the training and implementation phases. The time of planning and reviewing of each criterion and the implementation of the system of automatic editing, it was considered very short; due to the methodology, but above all to the experience of the staff of the central office.

33. In the same sense, the local office staff manifested that the time for the review of reports and delivery of informs was sufficient to achieve those responsibilities.

34. Respect to the strategy they considered that it was efficient, because the centralization of the criteria ensured that all them were applied in the correct order and with homogeneity for each

federal entities. The communication was effective; allocate an email account in the local offices to communicate every aspect related with the automatic editing meant an important achievement during the implementation; in addition to every responsible had the opportunity to communicate by telephone to central offices to resolve doubts immediately.

35. For this Institute, there is a need to explore and prove new methodologies for imputation and editing data as well as a deep analysis of the impact of the treatments, taking advantage of the wealth of information available.