

Design of “Quality Archives” for controls on Population and House Census Data.

Disegno dell’ Archivio di Qualità per il Controllo dei Dati del Censimento della Popolazione e delle Abitazioni

Angela Chieppa, Fernanda Panizon
ISTAT, Via Ravà 150, 00142 Roma (Italia)
chieppa@istat.it, panizon@istat.it

Riassunto: La sfida della qualità dei dati del Censimento della Popolazione del 2001, passa attraverso un ripensamento delle procedure di controllo e di validazione, che vengono strutturate, all’interno del “Sistema di produzione”, in un ben definito sottosistema di Controllo della Qualità dei Dati, realizzato in modo che sia possibile verificare e migliorare la qualità dei dati *durante* le elaborazioni. Tale Sistema di Controllo permette di registrare le informazioni sui controlli e sulle modifiche subite dai dati durante le elaborazioni: da tale “archivio” si possono estrarre a posteriori gli indicatori di qualità. Un prototipo di tale Sistema è stato realizzato per i dati provvisori del Censimento.

Keywords: Data Quality, Check and Editing, Population Census.

1. Introduction

To assure and improve the quality of Census data all the different operations, decisions, controls and actions have been considered as parts of the whole quality strategy. In order to reduce errors organisation has been planned, a monitoring system has been built to control external operations and fieldwork and a “System of data controls” has been designed to check and analyse the quality of data while they are processed, by the National Statistical Office (ISTAT), to eliminate systematic and stochastic errors and to produce accurate data and statistics for dissemination. The data processed and produced by this ‘Data Quality Control System’ are stored in a database, which can be considered both the archive of stored information about quality and a powerful instrument for improving data quality during data processing. The main objectives of such a system are: to enable comparisons with data from other sources, to enable exploratory analyses on processing data, offering a set of automated procedures for control, browsing and reporting on quality data, to produce a set of quality indicators for each processing step.

2. The Data Quality Control System

The design of the “Data Quality Control System” started by reconsidering the approach to controls of the previous Census. In 1991, control procedures were performed separately and lacked integration. Moreover, many controls based on comparisons with external sources of data, though planned, were not actually realised; only after disseminating Census data, some

analyses were carried out for producing quality indicators and for evaluating the accuracy of data. The new, more comprehensive approach implies that analyses and controls on data, measuring data modifications and collecting information about the different sources of error, have to be carried out while processing data. This approach of “continuous” controls, evaluating procedures' effects while they are running, allows us to improve data processing and, in the end, provides more information on quality, useful for final users. These considerations have made evident two specific issues in designing the Data Quality Control System for 2001: the integration of controls, their effectiveness and timeliness; the integration between controls and the processing phase.

The first can be obtained through implementing a “Control System” which involves:

1. a specific organization of human resources (including: revisers, running automated controls and checking their outputs, so assuring a first step of check on data; supervisors, coordinating groups of revisers, responsible for a second step of validation; analysers, responsible for exploratory analyses on data, assuring flexibility of control);
2. a specific database for quality data (“Quality Archives”), containing all the data required for controls;
3. a “library” of automated controls to be run on data;
4. specific applications to access/browse control data, to get better runtime reports, to notify inconsistencies or "suspicious" situations in data, and tools to analyse data “out of the usual processing line”.

The integration of the controls and processing phase, not only to assure the execution of every planned control, but also to allow analyses and checks not prepared in advance, is obtained designing the “Data Quality Control System” as a part of the more general Census Data Processing System; this implies that:

1. the “Quality Archives” are part of the Census database, so allowing navigation among micro-data being processed and their related “quality information” (that is a certain set of collected data can be combined, analysed and compared with previous information gathered on its own quality).
2. the automated controls are precisely scheduled among other procedures during the processing phase, that is, while data are being processed, planned controls on data are executed and in case of errors the following processing procedures could be stopped until the inconsistency is solved.
3. easy and "efficient" analyses on the data being processed are possible. The processing system is not operating rigidly, but it's possible to modify automated procedures on the basis of results coming from exploratory analysis. This means providing analysers with the necessary data (quality archives), to organise them to enable fast aggregation and to provide user friendly software tools to manage data.

3.The “Quality Archives”

To efficiently perform controls and analyses not only does all necessary data have to be collected in the “Quality Archives”, but also tables have to be created in which to store information on data modifications. The conceptual design of this quality database implies that some important areas have to be prepared and organised:

1. Meta-data and macro-data areas, related to the Census micro-data collected. It is essential to load meta-data tables (labels, definitions, formats, addresses of the tables, etc.) to allow exploratory and interactive analyses, as well as preparing pre-arranged tables, designed to ease queries and to get some results faster, such as aggregations by territorial level.
2. At the starting point of the processing phase, the archive will be loaded with data from other sources and information on data quality gathered in previous phases. Taking into consideration the various subjects of the Census (population, occupation, education, houses, buildings), the external data sources are mainly official statistics or data from surveys on the same matters. Many demographic indicators are based on the two annual population surveys carried out by ISTAT on the population registers situations; other information is taken using the Labour Force survey and surveys upon Education; others from surveys dealing with tourism; finally, aggregations on the previous Census data are considered to make significant comparisons. Moreover to have a complete “audit trail” of data, specific quality indicators are chosen and calculated to summarize the information collected during previous steps, particularly while monitoring the fieldwork and during the data capturing phase. These indicators could allow comparisons and investigations on the causes of errors.
3. Information about every kind of control carried out and their outcomes: rules and/or indexes used to execute controls/comparisons, thresholds of the alarms designed, and so on.
4. Data about processing execution: information on running procedures and about users of the systems, but chiefly information about modifications of data. This means that the quality archive will “grow” after every step of control, adding information about the modifications passed on data.

At the end of the processing step, all data from the Quality Archives will be used to produce indicators - both analytical and synthetic - describing data and procedures’ quality, for each step or unit.

4. Test of the Data Quality Control System with the provisional data

During the fieldwork operations, the 8101 Census Municipal Offices have to fill in some “auxiliary forms”, with data for each enumeration area and for different units of analysis (buildings, dwellings, households, institutional households, individuals). These provisional data (37 items) at Municipal level are transcribed in a specific “Provisional Data Form” to be faxed or e-mailed to the National Statistical Office. Here each item form is accurately checked and validated before dissemination. During this early phase an important preliminary test on the Data Quality Control System has been performed:

1. the organization of revisers/supervisors/analizers has been built up;
2. the Quality Archives have been loaded with data from external sources; a set of indicators for each Municipality have been chosen and organised in the database: indicators of “risk” (demographic size, presence of natural disasters in the last decade, the low degree of collaboration with Statistical Offices and the low quality of current population statistics) and indicators on demographic characteristics (ageing, migration), geographical typology.

Also important information from the monitoring system of the fieldwork has been loaded, such as number of questionnaires already revised by Municipalities, number of questionnaires already compared with the population register, and so on;

3. specific controls have been designed to detect errors/anomalies and an application has been developed to manage controls, their outcomes and to monitor the work in progress. Using this application, revisers can view, through forms, the stored data of each Municipality and ask for two different levels of controls on them. A first level of “formal control” deals with the lack of data in some cells of the form or with the inconsistency among cells which are correlated (sum of other cells, subtotals). A second level of qualitative control relied on external data (previous Census, official surveys and estimates), to verify the reliability of the data sent. For each data one or more comparison with external data are calculated and the differences are evaluated on the basis of a set of prearranged thresholds (low, medium, high). If some thresholds are overcome, the data have to be analysed better.

Considering the output of controls in the application, the revisers decided when the Census Municipal Offices had to be called (eventually rectifying anomalous data), when they can directly validate the data or to pass the inconsistency to the supervisors/analysers for a deeper analysis. To allow this kind of study in depth a specific environment has been set up for querying the database. During the process of data validation, the quality archives grow, with information about executed controls, errors, modifications. At the end of this step, the quality archives will be updated by summarizing information previously gathered: which revisers controlled the data, reference dates (arrival, contact with the Office, validation), errors, modification, the stage reached on work at the Municipal Offices when they sent the provisional data (to have an evaluation of the “accuracy/reliability” of the provisional data), and so on; in this way enlarging once more the quality archive. Testing the Data Quality Control System on provisional data has implied not only an improvement of quality of the initial coming Census data, but also an important training experience, helping us reconsider practical and working aspects. To realise the Data Quality Control System for Census data means actually making available a powerful instrument, to win the challenge of data quality assurance.

References

- Colombo, B. (1995) *La qualità dei dati dell'ultimo Censimento generale della popolazione e delle abitazioni*, paper prepared for "Commissione per la garanzia dell'informazione statistica".
- Egidi, V. (1999) Le strategie dell'Istat per i Censimenti del 2000, in “*SIS: Verso i censimenti del 2000*”, Udine.
- Engstrom, P., Granquist, L. (1999) Improving Quality by Modern Editing, paper presented at: *UN/ECE: Work Session on Statistical Data Editing*, Roma.
- Filippucci, C. (2000): Qualità delle statistiche e controllo del processo di misura, paper presented at: *SIEDS-ISTAT: La qualità dell'informazione statistica*, Roma.