**Data collection quality control using paradata and geolocation**

Marcos Rodrigues Pinto, IBGE, marcos.pinto@ibge.gov.br

**Abstract**

*This paper describes the progress made by the Brazilian Official Statistics Office (IBGE) to assure quality on its latest census. Management information systems are being extensively improved to deliver critical information to all the levels of administration on the data collection phase, producing important data for decision making. Tools like enumerator tracking by geolocation brings new possibilities to field monitoring, as it makes available the enumeration area coverage, not seen solely with the household coordinates registered before an interview on a CAPI approach. Also, the use of paradata (data about the process of data collection, produced by the enumerator) has shown it´s extremely promising value on quality control and consequential cost saving. The behaviour of the enumerator on the use of the handheld device can give detailed information concerning quality and productiveness. Examples of data that are currently used in Brazil are the record of coordinates during questionnaire execution, enumerator navigation on the questionnaire, interview length and questions sequence. Analysing this data while the collection is in progress and sharing with the field supervision has been shown very important as they can act preventively, checking an enumerator work that is signed as suspicious (possibly with frauded questionnaires), and making correct decisions. This work shows the potential use of paradata and geolocation as a fundamental piece of management information systems to produce high quality official statistics*.

**Keywords:** Paradata, data collection, geolocation, management information systems

1. **Introduction**

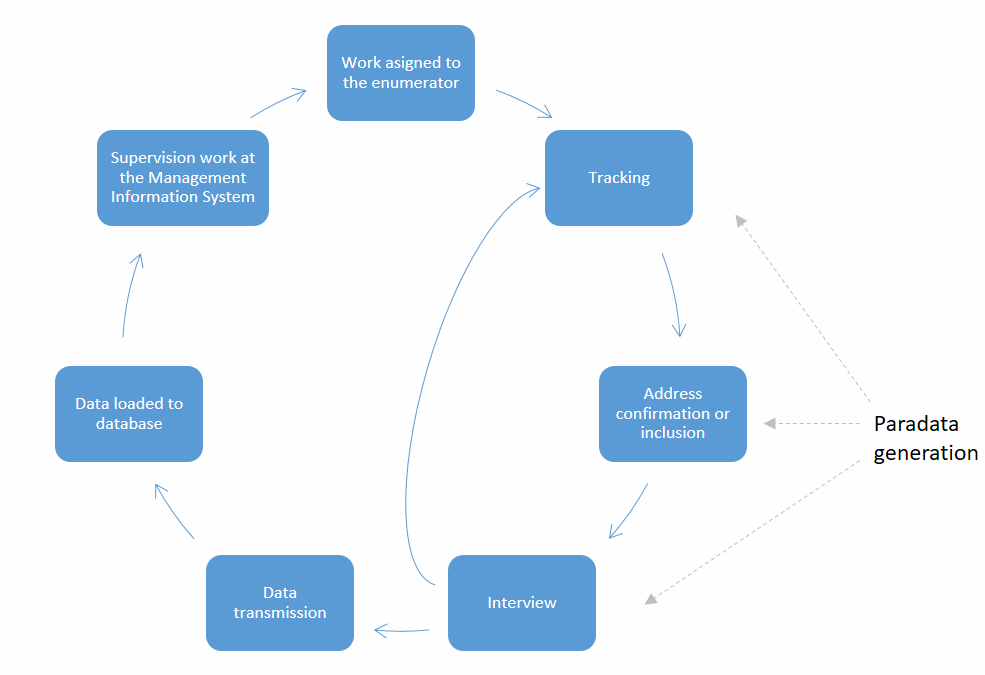
Quality control in the census data collection phase is vital to produce quality statistics. The field work performed by previously planned guidelines will certainly increase the likelihood of obtaining representative statistics on the quality scale. However, inaccurately or fraudulently captured data may lead to undesired distortions, which must be identified as soon as possible for necessary corrective action to be taken, specially while the field operation is still on progress. It should be stressed that a major goal of any quality assurance programme is to detect errors so that remedial actions can be taken even as the census operations continue (UNSTATS, 2017). The most common way for quality control consists in verifying and analysing the questionnaire data that is being obtained, as the collection advances in the field. The generation of indicators can qualitatively represent these data, which can be evaluated through a management information system (MIS). Besides that, another important and promising way is the use of paradata, that is discussed in this paper.

Despite its widespread use, the term “paradata” was never formally defined. Its more narrow conception as automated data about the process of interviewing has grown to encompass a variety of other types and sources of data (Couper P, 2017). The term has expanded to cover all types of data about the process of collecting survey data such as interviewer call records, length of interview, keystroke data, interviewer characteristics (Nicolaas G, 2011). The use of paradata in the context of CAPI data collection mode becomes increasingly viable as modern surveys tends to use devices that have touch screen, high storage capacity and in-built GPS. Thus, the capacity to store a large amount and variety of data, different from the questionnaire data, is increased, even making the volume of paradata much greater than the microdata. Paradata can play an important role in the application of continuous quality improvement methods (Kreuter, 2010). This resource has been increasingly effective and complementary to the analysis of the microdata, as was the case of the Brazilian census of agriculture of 2017, where the analysis of the behaviour of the enumerators on the field showed up to be very efficient for supervising control.

1. **The data collection in the Brazilian census of agriculture of 2017**

The Brazilian census of agriculture of 2017 have collected more than 5 million questionnaires in the 127 thousand enumeration areas, during a few months period. This census involved the hiring of more than 22.000 temporary workers, who assumed the role of 18.000 enumerators and 4.000 thousand field supervisors. he whole staff equipped a smartphone as a working tool for conducting a questionnaire interview with hundreds of questions. The collection process starts with the work assignment, where an enumeration area is allocated to the enumerator, and its device is loaded with satellite image and also a list of addresses corresponding to its assignment. When performing the enumeration area tracking, the enumerator is responsible for visiting the households present in the address list and also include the new ones. When performing confirmation of an address or adding a new one, the GPS is triggered to record the coordinates of the site, a process that is perceived by the enumerator on the device application. Then a questionnaire can be started. At the end of the day, or during it, it was recommended to make data transmissions to the central servers, step that can be done in any place with internet connectivity. The transmissions made during the day were accumulated on the central servers and loaded into a central database daily. These data were processed to generate several indicators that provide the work of the supervisors, who analysed them in different ways in the management information system (see Figure 1). This system was responsible for the field work follow-up in a qualitative and quantitative manner by using tables, charts, cartograms and many reports.

**Figure 1. Data collection workflow**



1. **The paradata generated in the Brazilian census of agriculture of 2017**

The paradata generated during the visit of enumerator encompass two groups of data: geocoordinates and data about the behaviour of the interview. An example of the first group was the enumerator tracking on the field. This data was captured automatically by the devices while the enumerator went through the enumeration areas by triggering the GPS each displacement of 16 meters. The capture of this data had as motivation the analysis of the enumerator trajectory and verification of the coverage of the enumeration area. Another type of geocoordinate was obtained during the interview by triggering the GPS every two minutes. This action is performed without the enumerator being notified of the fact, thus allowing a subsequent work of comparative analysis of the proximity between the coordinate captured in the address confirmation (which has visibility of the enumerator) and the coordinates captured silently during the interview.

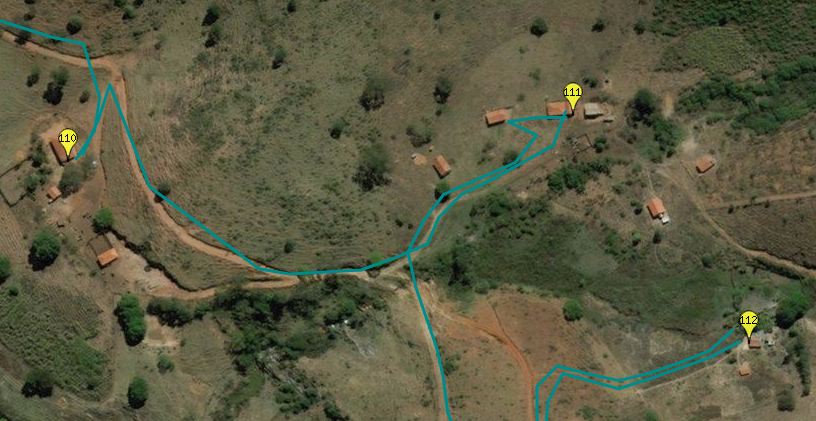
The data generated on the completion of questionnaire comprises a complete log of information captured from the start to the end of an interview, representing all actions taken by the enumerator in his handheld device. This information contains a time record and identification of the enumerator on each navigation action of the questionnaire (next button press, back button press), answer modification, time taken to answer each question, questionnaire reopening, question navigation and interview length.

The volume of the generated data is very expressive and represents the great extent of the Brazilian censuses operations. Before the end of the data collection, there were more than 800.000.000 GPS tracking coordinates, more than 90.000.000 GPS questionnaire coordinates and more than 2.000.000.000 registers of questionnaire navigation.

1. **The paradata usage during the data collection**

The paradata captured during the data collection on each mobile device were constantly transmitted to the central datacentre and retained for analysis by the supervisors. The coordinate capture was fundamentally used to monitor the collection coverage, visually allowing to identify if the enumerator tracked the entire enumeration area. This monitoring was possible by plotting the geographical coordinates of these points in a linked way, hence representing a tracking of the course (see Figure 2 and Figure 3).

**Figure 2. Enumerator tracking in a rural area**



Source: IBGE

**Figure 3. Enumerator tracking in river areas**



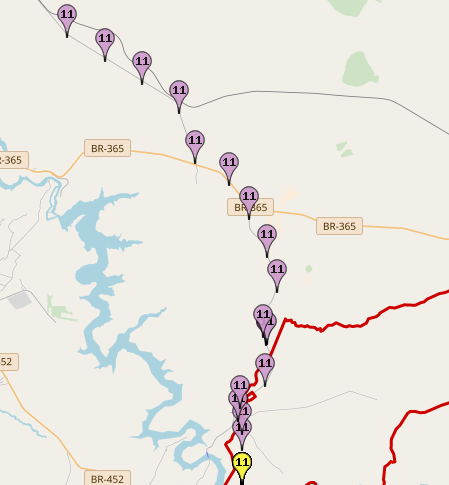
Source: IBGE

Moreover, the use of the questionnaire coordinates was presented as a great tool to verify possible attempts to interview in different places other than the expected address, as well as fraud attempts, where the enumerator completed a questionnaire without the presence of the respondent. As the payment of the census enumerators was carried out by productivity, frauded questionnaires could have happened. The interview coordinates represent a silent way to monitor the geolocation of the questionnaire completion and identify unexpected behaviours, such as:

* Interviews made far from the capture of the address coordinate
* Interviews held in motion
* Many interviews conducted in the same place

Each of the indicators shown above were used in the MIS by geolocating this information, and let the supervisor adjust the parameters he considered suitable (ie: view distant interviews more than 300 meters from the expected address) and by this way, identify the suspicious cases (see Figure 4 and Figure 5). The use of this resources allowed the field supervisors to make corrections during this data collection phase, so making possible to request re-interviews before the end of the enumeration area work.

**Figure 4. Interview held in motion and outside the enumeration area**



Source: IBGE

**Figure 5. Many interviews conducted in same place**



Source: IBGE

1. **Conclusion and future research**

This paper demonstrated the relevance of use of paradata monitoring the field work of the 2017 census of agriculture census, by complementing the analysis of the microdata and becoming a powerful tool for measuring the quality of the data collection phase. The data generated by the enumerators handheld devices allowed the field supervisors to check the quality of their work, by analysing tracking, questionnaire coordinates and interview length, making possible to identify suspicious or fraudulent work.

Vast amount of data related to questionnaire navigation was capture automatically by the handheld device. However, these navigation data were not totally explored in this census. Further analysis may disclose valuable information regarding fraudulent patterns in this data. In this regard, a machine learning model can be possibly designed to check the quality of the collected data by identifying the suspicious questionnaires automatically in future census operations.

1. **References**

UNSTATS, (2017). Principles and Recommendations for Population and Housing Censuses: Revision 3 (Statistical papers). United Nations. pp 66

Couper,Mick P. (2017). Birth and Diffusion of the Concept of Paradata (in Japanese – Translated by W. Matsumoto). Advances in Social Research, 18, 14-26.

Kreuter, F., Couper, M., Lyberg, L., (2010). The use of paradata to monitor and manage survey data collection. pp 285-287

Nicolaas, G. (2011). Survey Paradata: A Review. National Centre for Social Research (NatCen), January 2011. Accessed 10 May 2018. Available at: <http://eprints.ncrm.ac.uk/1719/1/Nicolaas\_review\_paper\_jan11.pdf>