WP. 12
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (ii): Editing near the source

EDITING MULTIMODE DATA COLLECTIONS: THE SPANISH EXPERIENCE

Supporting Paper

Prepared by Pedro Revilla, Ignacio Arbués and Soledad Saldaña, INE, Spain

I. INTRODUCTION

1 Traditionally in INE data were collected by enumerators or through self-administered mail questionnaires. Nowadays, Electronic Data Reporting (EDR) methods offer new opportunities for getting high quality incoming data reducing costs at the same time. Hence, like many others statistical agencies, INE has a significant interest in EDR methods, in particular in Web-based data reporting. An example of this was the possibility offered to all citizens to fill in the Population Census 2001 using the Internet. Concerning business surveys, Web questionnaires are offered as a voluntary option. A major target of this action is offering the respondents another option to fill in the questionnaires, in the hope of reducing respondent burden, or, at least, improving our relationship with them.

2 Nevertheless, for most surveys, EDR cannot be at the moment the only way of data collection. Paper data collection is probably going to stay with us for some years. Hence, a mixed mode of data collection (partly paper, partly electronic) should be used. In other cases, multimode data collection is used as a way to reduce survey costs.

3 Multimode data collections have a strong impact on data editing strategies, because data editing may differ depending on the channel used. Global strategies should be designed in order to guarantee that similar edits are applied to all the questionnaires avoiding mode effects.

4 The Spanish 2009 Agricultural Census will use a multimode data collection via four channels. The different channels will be organized in the collection schedule from the cheapest to the most expensive: mail, CATI and PAPI. The CAWI option is also open during the entire collection time span.

5 The Census data editing strategy follows two key ideas. First, the same set of edits has to be pass for every questionnaire, whichever channel it comes, trying to avoid mode effects. Secondly, a selective editing strategy has to be used to manage such a big set of data.

6 In order to design a selective editing method, we have carried out an experiment using the microdata of the former Census. The selective editing method is based in stochastic optimization techniques. We will determine a selection strategy that allows editing the minimum number of units, while obtaining certain accuracy requirements in the aggregates. The selection strategy consists of a score function that is built as a linear combination of the conditional expected quadratic errors of each variable of the questionnaire.

7 The remaining of the paper is organized as follows. Section II discusses the challenges and opportunities of editing multimode data collections. Section III presents the data editing procedures planned for the 2009 Agricultural Census. The paper ends with some final remarks.

II. CHALLENGES AND OPPORTUNITIES OF EDITING MULTIMODE DATA COLLECTIONS

8 When designing a survey the target is to optimize the production process in order to reduce total survey error within the available budget. In many examples multimode data collection may be the best affordable method. In many cases, the use of multimode data collections by INE has been the straightforward consequence of our policy of empathizing EDR reporting methods on one hand, but leaving to the respondents the decision to choice the reporting channel on the other hand. In other examples multimode data collections have been the consequence of choosing the most cost-effective method.

9 In both situations, multimode data collection allows benefiting, even partially, from the advantages of EDR. Whereas Computer Assisted Interviewing integrates into one stage previously distinct phases such as interviewing, data capture and editing, Computerized Self-Administered Questionnaires (CSAQs) go a step further by shifting such activities to the respondent.

10 Several advantages could be expected from using CSAQs questionnaires. Improving accuracy results from built-in edits, which allow the reporting enterprises to avoid errors as they are made. The elimination of data keying at the statistical agency directly gets rid of a common source of error. Some electronic devices (automatic data fills and calculations, automatic skipping of no applicable questions, etc.) could help the respondent to fill in the questionnaire easier and faster. The cost for statistical offices to carry out a survey using the CSAQs questionnaires could decrease. Savings could be achieved from reducing storage, packing, postal charges and eliminating data keying and keying verification. Some of the editing tasks could be reduced from built-in edits. Finally, audit trails can be recorded from the electronic questionnaires and used in usability studies

11 Concerning the edits to be implemented, some crucial questions arise: What kind of edits should be implemented on the CSAQs questionnaires? How many? Only fatal edits or fatal edits and query edits? What kind of edits should be mandatory? On one hand, we need to include some edits. If we do not, then the information collected by CSAQs questionnaires should be treated to the editing procedures in exactly the same way as collected by paper. In that case, we would lose an essential advantage of CSAQs questionnaires: no need to editing again the information with a suitable set of edits implemented in the CSAQs application. On the other hand, we need to be extremely careful in the set of edits to be implemented, because if we implement a big set, then respondents will give up and prefer the freedom they have in paper. Too many edits could even irritate the reporting enterprises and increase the burden. In that case we will lose all the advantages of CSAQs questionnaires, as users will prefer the easy way (paper).

12 Many statistical offices are experimenting with the use of different EDR options in data collection (CATI, CAPI, XBRL, etc.). Web questionnaires offer some advantages over other more complex EDR methods. The Web is a mature technology for EDR because of widespread public acceptance in enterprises and institutions. The prerequisites are only a PC, access to the Internet, and a browser. There is no need, in principle, to incorporate other software on the reporting enterprises. The Web makes it simple to put electronic forms at the disposal of almost every enterprise, whatever its size.

13 Because the reduction in the enterprise burden using Web questionnaires is not always obvious, we try to encourage the use of Web questionnaires. A key success factor in encouraging the use of Web questionnaires is giving enterprises some incentives (temporary access to information, free deliveries of tailored data, etc.). Working in this direction is the last step in our TQM project of giving reporting enterprises free of charge tailored data in exchange for the questionnaires (Arbués et al., 2006). When an enterprise sends a valid form (i.e. passing the mandatory edits), it immediately receives tailored data from the server. These tailored data consist of tables and graphs showing the enterprise trend and its position in

relation with its sector. Offering this data through the Web has some advantages (speed, possibility to edit the file) over sending this same data on paper by mail. Taking these advantages into account, we expect more enterprises to use the Web survey. We are testing this action in the Turnover and New Orders Survey. This monthly survey uses a very simple form that includes only two variables: turnover and new orders, broken down by geographic markets.

III. A STUDY CASE: THE 2009 SPANISH AGRICULTURAL CENSUS

A. Census methodology overview and data editing strategy

14 INE carries out an agricultural census every ten years. Since Spain joined the EU in 1986 the Agricultural Census follows the EU legal and methodological framework. Up to now the agricultural censuses have been paper based operations and the data collection has been carried out using enumerators. A traditional methodology was used (i.e. all variables were collected from all population units, using the same collecting method, at the same time).

15 INE identified a number of drivers for a significant change in the 2009 Agricultural Census. The extension and increasing quality of agricultural administrative registers provide a large set of information on holdings. These data, in connection with household data in INE registers, allow constructing an initial frame in a way that mail out deliveries of census questionnaires become a viable option. Moreover, this administrative data allow improving imputations or even estimating some variables. Another driver is the increasing acceptance of Internet technologies by both the citizens and the statistical office. In a closely related development, the government policy in electronic administration pushes the different departments to move towards ensuring that an electronic option is offered. And finally, the need for an increased efficiency due to severe budget cuts was the definitive driver for the most significant change in the agriculture census methodology since the first census carried out by INE in 1962.

16 As a consequence of the above-mentioned drivers the census collecting method have changed from a single channel (paper using enumerators) to a multimode process via four channels. The different channels will be organized in the collection schedule from the cheapest to the most expensive: mail, CATI and PAPI. The CAWI option is also open during the entire collection time span. The Agricultural Census 2009 data collection will be carried out by means of two different questionnaires: a general questionnaire with those variables that will be investigated exhaustively and a sample questionnaire with data on agricultural production methods.

17 The paper version of the census questionnaire consists of 8 pages, opposite to the 16 used in the 1999 Census. It has been tried to simplify it as much as possible, requesting only those strictly necessary variables and obtaining the other ones by means of other sources or deducing them from other variables in the questionnaire, when their nature allows it. Examples of this are the holder's age that will be obtained from the population register using the ID card, or the holder's juridical personality that will be deduced from the NIF code (fiscal identification number).

18 The questionnaire is structured in 17 blocks of questions that gather information about different aspects:

(a) Holder and holdings: census receiver identification and location; holder identification and number of their holdings (one or more than one)
(b) Livestock: number of heads by species and location
(c) Lands: surfaces of crops, fallow lands, kitchen gardens, permanent grassland and meadow and other lands; type of tenure and location of the lands; total irrigated area during the agricultural campaign; energy crops for the production of renewable energy and genetically modified crops
(d) Organic farming
(e) Labour force: holder and manager information and working days for each of them; family and non-family work and farm work undertaken on the holding by persons not employed directly by the holder
(f) Equipment used for renewable energy production by type of energy source

(g) Rural development.

In addition, the sample questionnaire has 4 pages with 8 sections that contain information about production methods on the holding. It provides data on:

(h) Production methods on the holdings with livestock: animal housing; animal grazing and manure storage and treatment facilities;
(i) Production methods on the holdings with lands: landscape features; soil conservation; tillage method and manure application;
(j) Complementary variables on irrigation: average irrigated area the last three years; irrigation methods employed and source of irrigation water used on the holding;
(k) Information about other gainful activities of the holding directly related to it done by the holder, the manager and family and non-family work.

19 Editing an agricultural census is a quite difficult issue. The large number of units to edit and the length and complexity of the questionnaire make implementing an efficient editing strategy and an appropriate set of edits a cumbersome task. The methodological changes in the 2009 Census (for example the use of two different models of questionnaires) and the multimode data collection process put new challenges on the data editing and imputation phase. In addition, the much smaller labor force gives rise to the need for a system that makes more editing and imputation decisions automatically, without manual intervention.

20 Two ideas were clear from the beginning of the project. First of all, the same set of edits has to be passed by every questionnaire, whichever channel it comes, in order to avoid mode effects. Secondly, a selective editing strategy has to be used, in order to manage such a big set of data.

B. Editing during data collection

21 The subject matter team has designed a set of deterministic edits that will guarantee the internal consistency of information collected in both questionnaires, as well as the consistency between them. The above-mentioned edits consist of a list of 49 edits: 23 for the census questionnaire and 26 for the production methods questionnaire and its relations with the census ones. The edits are organized by module, where a module corresponds to a section of the questionnaire.

22 These edits detect errors such as lack of response in questions where the flow of the questionnaire require it (for example, holders say that they are owners of a holding but there is no data on livestock or lands in the questionnaire), inconsistencies in the information of crops and cattle indicated in different blocks of the questionnaire (for example, cereals hectares from organic farming is larger than the total cereal area of the holding), or values out of range (for example, the holder work is more than 365 working days).

23 Besides the former edits, there is a set of checks for the paper questionnaires in order to split questionnaires that will have to go through telephone follow up to obtain some additional breakdown. In particular, the questionnaire requests both for family and non-family labour regularly employed on the holding, the total of persons and working days worked by gender. Since in the final file these data must figure in an individual level, when the number of persons in one of these groups is greater than one, it will be necessary to obtain a breakdown of the above-mentioned data.

24 Every questionnaire will be checked with these edits whichever channel the questionnaire has been received. Paper questionnaires will be scanned and captured. Data capture will be carried out on the scanned image, using Optical Recognition of Marks (OMR). By means of this technology, the recording application will focus just on those fields that have information in the questionnaire. Moreover, this process will be developed in a specialized way according to the type of information that will be recorded (literal, qualitative or quantitative data) and according to the section of the questionnaire. Two independent recordings will be obtained, with a third one used as a tie-breaker in case there is discrepancy between the first two ones. After that, errors will be detected and questionnaires with inconsistencies will be follow-up by telephone, when telephone is available. We will get in touch with the

informants again to solve the errors in their questionnaires. Telephone calls will be carried out according to some criteria of priority relation with the holdings.

25 The Computer-Assisted Web Interview (CAWI) system has been programmed so that wrong data is detected automatically when the respondent is filling in the questionnaire. When the application finds an error, it will show an error message warning of it. The informant could solve the problems immediately or continue filling the form without doing it. At the end of the questionnaire, a list with uncorrected errors will be shown and it will be possible again to solve them. The system will allow sending questionnaires without correcting if they comply with certain requirements of information mainly related to data on holder identification and location and lands and/or livestock. Questionnaires received by web with errors will be following up by telephone in the same way than paper ones.

26 When data are collected by CATI, the application will direct the flow of the questionnaire and ensure the fulfilment of the rules. The system will warn when an error is detected and interviewers will be able to contrast immediately any problem of inconsistent data with the respondent. Hence, this questionnaire will be clean when the interview ends. Finally, paper and pencil interview will be carried out by qualified staff that will check the consistency of data supplied by informant at the moment of the interview.

27 As it has been mentioned, everything above corresponds to a editing process whose aim is to guarantee the internal consistency of the information provided by every holding. Later, when data from all channels are arranged and joined in a unique file, they will be submitted for further editing processes. In this phase, we expect selective editing to play a central role.

C. Selective editing

28 In order to design a selective editing method, we have carried out an experiment using the microdata of the 1999 census. These data can be downloaded from the web of INE .

29 Data from about 1.75 million farms were collected in the census 1999. The questionnaire comprises near 300 variables, but given that the questionnaire of the new census was intended to be simpler, we selected a subset of around 250 variables.

30 The selective editing method is based in the techniques described in Arbués et al. (2009). We will determine a selection strategy that allows editing the minimum number of units, while obtaining certain accuracy requirements in the aggregates. The selection strategy consists of a score function that is built as a linear combination of the conditional expected quadratic errors of each variable of the questionnaire.

$$\delta_i = \sum_{i=1}^{k} \lambda_j E\left[\varepsilon_{ij}{}^2 \mid \Im\right],$$

31 where i is the index of the questionnaire, *j* the index of the variable, ε_{ij}, the error, λ_j some coefficients and $\Im$ is the σ-field that comprises all the available information. We compute the conditional expectation of the error using the formulas of Arbués et al. (2009). These calculations involve the following:

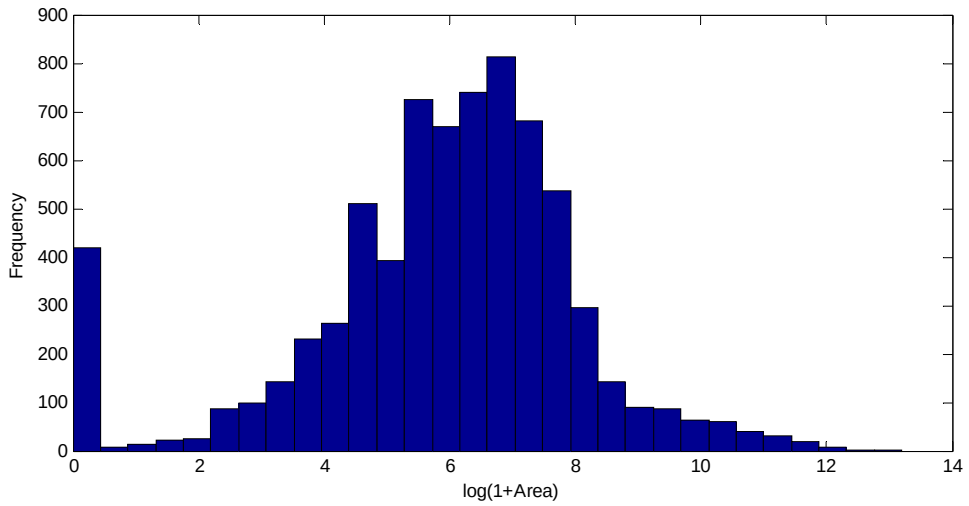
(a) The probability of measurement error.
(b) The variance of the measurement error.
(c) A prediction $\hat{x}_{ij}$ of the true value of each variable x_{ij} and the standard deviation of the prediction error.

32 The first two values have to be set *a priori*, but they can be calibrated to take into account the differences in quality of the data gathered using various modes of collection.

33 In previous applications of this methodology to short-term indicators, we had calculated the prediction using simple time series models. In this case, we do not have time series, so we need a static model, one that exploits the simultaneous relationships between the variables of the questionnaire.

34 Most of the variables are quantitative (land area, livestock, etc) — only twelve are categorical. The quantitative variables are strongly skewed and after a logarithm transform, most of them have distributions that resemble a mixture of a degenerate X=0 variable and a normal (see in figure 1 an example). They also present frequency peaks at certain round numbers (e.g., 100a). There are a few variables with continuous bimodal distributions, which we did not try to model specifically.

Figure 1: histogram of the distribution of the variable "land in property", transformed.



35 Our aim is to predict the transformed variable $y_{ij} = \log(1 + x_{ij})$ using the other variables $y_{ik}, k \neq j$. We tried at first to use a logit or probit regression to distinguish the $y_{ij} = 0$ case from the normal-distributed one and then, to predict the latter by linear regression. Unfortunately, the discrete model has presented serious difficulties, both in the selection of a set of regressors among the large set of possible ones and also in the estimation of the parameters. Consequently, we do not try to predict the case $y_{ij} = 0$ as a special case and we just use a linear regression model,

$$y_{ij} = \beta_1 z_{i1} + \ldots + \beta_p z_{ip} + \xi_{ij}$$

36 where $(z_{i1}, \ldots, z_{ip})$ is a vector of regressors selected among the full vector of variables $(y_{i1}, \ldots, y_{ik})$ and dummy variables for the categorical ones.

37 The predictive ability of the linear model varies greatly among the variables. The worst cases are mostly found among those variables whose distribution is less informative because they have very few nonzero values (hereafter, FNZ variables). Let us measure the quality of the regression by its R^2. In the upper plot of figure 2 we present a histogram of the R^2 of all the variables. If we weight the variables according to the number of nonzero values, we see (lower plot) that the quality of the regression is generally good for the variables with more nonzero values.

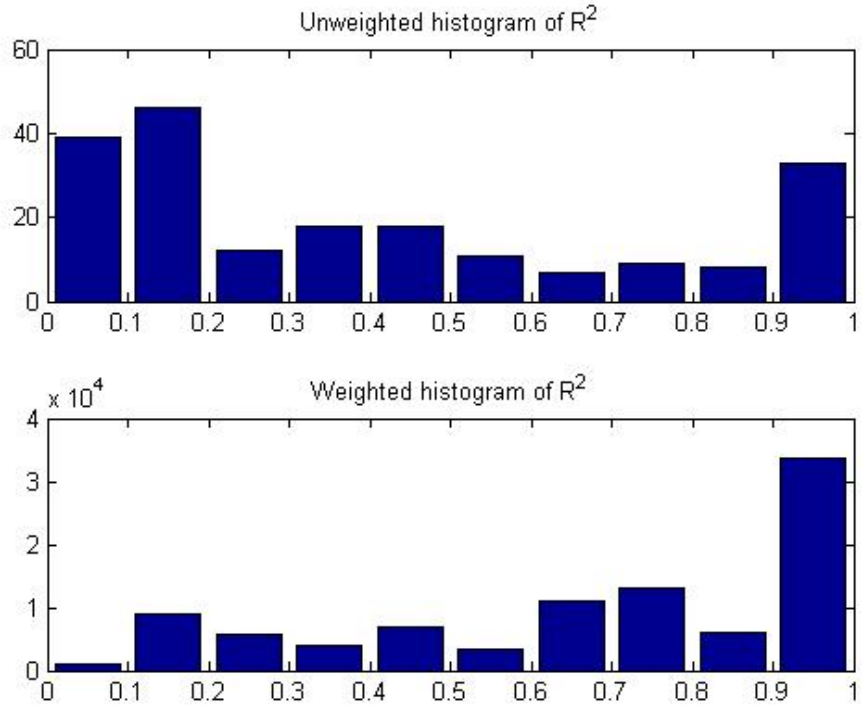


Figure 2: Histograms of the R^2 of the linear regressions, unweighted and weighted.

38 The consequence of this is that the FNZ variables are predicted with a large uncertainty. On the other hand, those few units represent a large share of the total of the FNZ variables. Consequently, our method tends to select many questionnaires among the ones that have those few nonzero values.

39 The prediction of the categorical values presents the same difficulties as the prediction of the zero values of the numerical ones, so the selective editing is for the moment based only in the quantitative variables. A possibility worth to be investigated in the following stages of our work is to use principal components to reduce the number of regressors both for the quantitative and the qualitative variables.

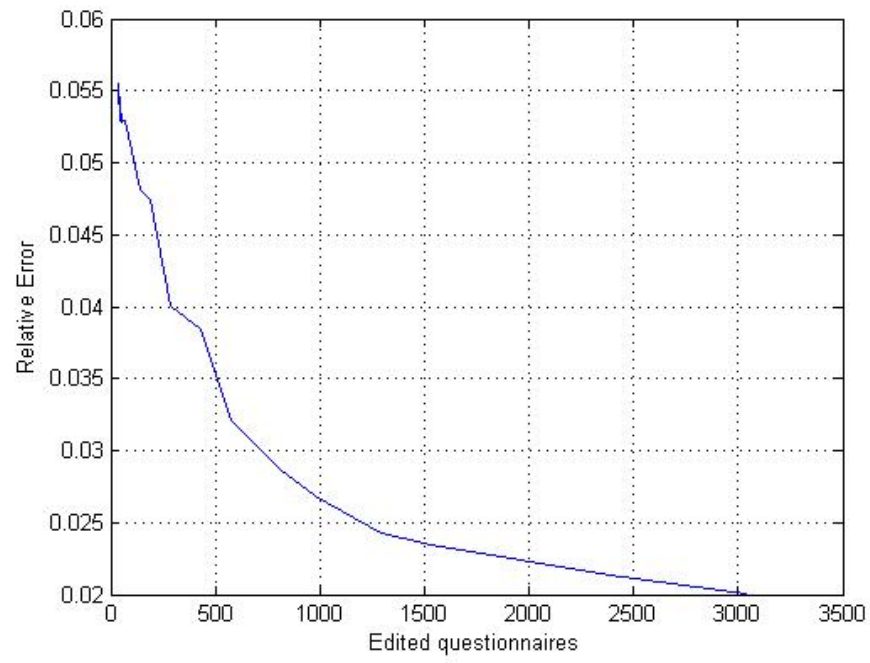


Figure 3: Remaining error as a function of the edited questionnaires (N~9000).

40 In order to evaluate the performance of the method, we have made a simulation, introducing random errors in the real microdata and trying to correct them using our selective editing approach. In order to avoid working with huge data sets, we make the selection and evaluate the aggregates at the province level (NUTS 3).

41 In figure 3, we show an example curve that represents the remaining relative error (average across variables and across simulations) as a function of the number of questionnaires edited. Compared with previous applications of our method, this curve decreases more slowly, not surprisingly, given that working with so a complex questionnaire, more units have to be edited to purge all variables from error.

IV. FINAL REMARKS

42 A prerequisite of any editing strategy is getting high quality incoming data. In some surveys mixing modes may be the best affordable method. Multimode data collections have a strong impact on data editing strategies, because data editing may differ depending on the channel used. Global strategies should be designed in order to guarantee that similar edits are applied to all the questionnaires avoiding mode effects.

V. REFERENCES

Arbués, I., González-Dávila, M., González, M., Quesada, J. and Revilla, P. (2006). Using a TQM approach to get high quality incoming data in the Spanish industrial surveys. European Conference in Quality in Survey Statistics. Cardiff. April 2006

Arbués, I., González, M., Revilla, P. (2009), A class of stochastic optimization problems with application to selective editing(working paper, it can be obtained from the authors).

INE Agricultural Census, 1999. URL: http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft01/p042&file=inebase&L=1