

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**  
(Bonn, Germany, 25-27 September 2006)

Topic (i): Editing nearer the source

**EDITING WEB QUESTIONNAIRES: CHALLENGES AND OPPORTUNITIES**

**Invited paper**

Submitted by National Statistical Institute (Spain)<sup>1</sup>

**I. INTRODUCTION**

1. This paper discusses the challenges and opportunities of Web questionnaires in order to improve editing tasks. Web questionnaires offer new possibilities on moving editing closer to respondents. Whereas Computer Assisted Interviewing (CAI) integrates into one step previously distinct phases such as interviewing, data capture and editing, Web surveys go a step further by shifting such activities to the respondent. Hence, Web surveys offer the opportunity for re-engineering editing processes, in a way reporting enterprises may play a more active role in data editing. The possibility of using built-in edits allows reporting enterprises to avoid errors as they are made. The elimination of data keying at the statistical agency directly removes a common source of error.

2. Web surveys offer some advantages over other more complex Electronic Data Reporting Methods (EDR). 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 make electronic forms available to almost every enterprise, whatever its size.

3. As many other statistical agencies, the INE has a significant interest 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 enterprise surveys, the INE has launched a general project of giving respondents the option of submitting their responses to statistical surveys using the Internet. A major target of this project 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. Hence, a mixed mode of data collection (partly paper, partly electronic) is used.

Global strategies should be designed, because data editing strategies may differ when using paper that when using an electronic questionnaire. We are carrying out some studies related to the data quality of the two kinds of sources and the data editing implemented in each case.

---

<sup>1</sup>Prepared by I. Arbués (iarbues@ine.es), M. González-Villa (mgonzalez@ine.es) and P. Revilla (previlla@ine.es).

4. The main focus of this paper is on exploring the possibilities of integrating a mixed mode of data collection into a selective editing approach. The combination of built-in edits and selective editing approach appears very promising. Our target for the future is that, after implementing correct Web edits, traditional microediting could be dramatically reduced. In order to design a global editing strategy we try to introduce a theoretical framework to develop score functions. Some practical experiences in the Spanish Monthly Turnover and New Orders Survey are presented in this paper.

5. The following section is a general discussion on the challenges and opportunities of Web questionnaires. In section III, the possibilities of integrating a mixed mode of data collection into a selective editing approach are presented. The paper ends with some final remarks.

## II. CHALLENGES AND OPPORTUNITIES OF WEB QUESTIONNAIRES: GENERAL ISSUES

6. Web questionnaires present new opportunities and challenges in order to improve editing tasks. In mail surveys using paper questionnaires data editing has typically been carried out after data collection. However, computer assisted methods allows to carry out data editing interactively, during data collection. Moving editing closer to respondent can significantly contribute to improve editing effectiveness. Whereas Computer Assisted Interviewing (CAI) integrates into one step previously distinct phases such as interviewing, data capture and editing, Web questionnaires go a step further by shifting such activities to the respondent. Hence, Web surveys offer the opportunity for re-engineering statistical production processes, in a way reporting enterprises may play a more active role in data editing and provide high quality data.

7. Many statistical offices are experimenting with the use of different EDR options in data collection. Web surveys 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.

8. Several advantages could be expected from using Web questionnaires. These include improving accuracy and timeliness, and reducing survey cost and enterprise burden. 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 removes a common source of error. Moreover, this elimination of data keying reduces the processing time of the survey. There are other factors that can also contribute to improve timeliness. Data transfer on the Web can be done much faster than using the postal system. Some electronic devices (automatic data fills and calculations, automatic skipping of no applicable questions, etc.) could help the respondent to fill in the questionnaire faster. The cost for statistical offices to carry out a survey using the Web could decrease. Savings could be obtained from reducing storage, packing, postal charges and eliminating data keying and keying verification. Some of the editing task could be reduced from built-in edits. Nevertheless, to get the target of reducing enterprise burden using Web questionnaires is not so straightforward. The reduction in the enterprise burden is not always obvious. The respondents' benefits depend largely on the way metadata support the respondent in filling in the questionnaire (help texts, auto-fill rules, pre-filled data, etc).

9. There are a lot of expectations about the role of Web surveys in the years to come. Nevertheless, the implementation of Web surveys has often been lower than expected. The take-up of electronic data reporting for statistical data by business providers is generally less than 10

10. More research is needed to look for the reasons why, up to now, the rate of using Web questionnaires is quite low, while technical requirements are available for many of the respondents. Probably, electronic forms have not the same advantages for the reporting enterprises than for the statistical offices. For many of the questionnaires, the most time consuming tasks are to look for the required data and computing the answers. There is no time difference between keying data on a screen and to fill in a questionnaire on paper. The advantages for the reporting enterprises would probably be bigger if the information could be extracted straight from their files. But this procedure may be expensive for both reporting enterprises and statistical agencies, because an initial investment is needed.

11. The respondents' benefits need to be clearly explained to convince them to use the Web questionnaire. An important element to improve the acceptance of Web forms among reporting enterprises is to consider Web questionnaires in a wider context of all their administrative duties and of all electronic data reporting. It is unlikely that reporting enterprises are willing to adapt their systems only for statistical purposes. Hence, statistical offices should be aware of the habits of respondents and try to adapt electronic questionnaires to these trends (for example, e-commerce, e-administration, etc.). A key success factor in encouraging the use of Web questionnaires is giving enterprises some incentives such as temporary access to information, free deliveries of tailored data (Arbués et al., 2006), etc.

12. In any case, for most of the surveys, Web questionnaires cannot be the only way of data collection. Paper data collection and associated procedures (like scanning) are probably going to stay with us for some years. Hence, a mixed mode of data collection (partly paper, partly electronic) should be used. Global strategies should be designed, because data editing strategies differ when using paper to an electronic questionnaire. Nevertheless, it is necessary that the edits are consistent across collections modes.

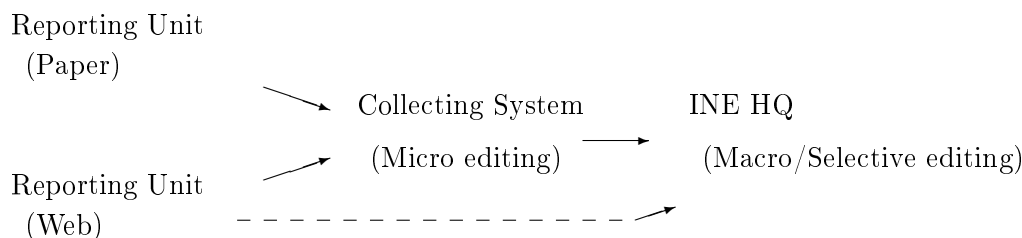
13. There are two contradictory targets implementing Web questionnaires. On one hand, to implement a single point of entry for all agency surveys, with a uniform security model and a common look across the entire site. And, on the other hand, to allow decentralised applications to cope surveys singularities. One aspect where the difference among surveys has to be taken into account is data editing. Combining the two targets (i.e. integrating a centralised platform with decentralised applications) is a non-trivial task.

14. Some crucial questions arise: What kind of edits should be implemented on the Web? How many? Only fatal edits or fatal edits and query edits? What kind of edits should be mandatory? When should the edits be performed? After each data item or after the whole form has been sent to the server? It is very difficult to devise rules to answer these questions.

15. On one hand, it is convenient to include some built-in edits. If we do not, then the information collected by a Web survey 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 Web surveys: the possibility to perform data editing interactively, during the data collection. On the other hand, we need to be extremely careful with the set of edits to be implemented in the Web survey, 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 Web surveys, as users will prefer the easy way (paper).

16. How to cope with the too few/too many edits dilemma? If we are trying to implement a Web questionnaire in an existing survey, a prerequisite is to analyse the current set of edits in order to determine the efficient set of edits to be used in the Web implementation. Monitoring the built-in edit

FIGURE 1. Collection/Editing Scheme.



function once the Web questionnaire has been launched is also a crucial point. Asking the respondents how they used the edit function and studying an edit log that records information each time the edit function is invoked (Weir, 2005) are suitable methods to evaluate it.

### III. INTEGRATING DIFFERENT MODES OF COLLECTION AND SELECTIVE EDITING

#### A. Description of the Survey

17. A monthly survey provides data to publish the Short Term Indicators of Industrial New Orders and Turnover. The sample size is of about 13,500 local units, for which 14 variables are requested:

- New Orders: Total new orders, domestic market, eurozone non-domestic, European Union non-eurozone, non-EU, stock at the beginning of the month, cancelled orders, orders invoiced and stock at the end of the month.
- Turnover: Total turnover, domestic market, eurozone non-domestic, European Union non-eurozone, non-EU.

The variables Total Turnover and Total New Orders, are the main target of the editing process, since they are used for the computation of the indices. When in the future, new indices are calculated, e.g. for the different markets, more variables will have the same consideration.

18. The structure of the data collection and editing of the New Orders/Turnover Survey is depicted in figure 1. There is a first micro-editing integrated in the recording of the data from the paper. Then, the data are transmitted to the INE headquarters, where more sophisticated macro and selective editing methods are applied. In the first phase there are both hard and soft edits. The hard edits check consistency relationships within the questionnaire and the soft ones are historical checks (ratio edits).

#### B. Integration of the web questionnaire

From January 2005 to February 2006, an experimental web reporting system was used to collect the data of the New Orders/Turnover Survey of around 80 enterprises. In February 2006, the INE offered a web questionnaire to all the enterprises in the sample. Among the roughly 13500 units sample, about 800 reported the data using the web the first month and this number has increased to over 1300 in

five months. The data inputted through the web is introduced into the same data stream of the paper questionnaires and then suffers the same treatment, including the first micro-editing (the edits in the computer application used to record the paper questionnaires are also applied to the web data). In the web questionnaire itself, there are some very soft edits (there are no hard edits and the user does not have to read the informative ones, but he or she has to click a link to see them).

19. There is another channel through which the raw data from the web system is sent to the INE headquarters without any micro-editing. This channel is much faster than the usual one, almost *real-time* (no more than 24 hours from the time the questionnaire is filled online). We want to analyze the possibility of using the data from the web to gain timeliness, reduce workload and avoid over-editing.

20. We can compare the raw web data with the same data after it goes through the collection/micro-editing system, and we have learnt that the data are extensively modified (28.9% of the questionnaires have at least one variable changed). Among them, 11.4% have some errors that correspond to *hard* edits, 12% lack some of the variables (many enterprises fill the turnover cells but not orders) and 5.5% correspond to other cases (filling the wrong cell, unit errors, ...).

Despite these many corrections, the changes in the main variables, measured in their influence on the published indexes is concentrated in not so many questionnaires, as we have checked in our assessment of the selective editing method.

An interesting fact is that the data received through the web system, is less frequently modified in the selective/macro-editing phase (2.4% of the web questionnaires are modified vs. 6.4% of the whole sample). There is not a significant difference of size between the enterprises responding through the web and the others that could explain this difference.

### C. Selective editing

21. The selective editing methods usually define a function (score function) that allows to select a certain number of data items to be manually edited. A good strategy would typically permit that editing a small number of questionnaires removes the most of the error (measured in terms of impact on the disseminated statistic). The current methods of selective editing are mostly empiric. In our work, we have tried to introduce a theoretical framework to guide in defining adequate score functions.

22. Let us define the problem of data selection for editing in a formal way.

Let us assume that we have  $N$  observations of a variable and we will use them to compute certain statistic. We will denote as  $x_t^i$  the true value and  $\tilde{x}_t^i$  the observed one. Thus, we write  $\tilde{x}_t^i = x_t^i + \varepsilon_t^i$ , where  $\varepsilon_t^i$  is the observation error. Also, the theoretical unobserved statistic is  $I = f(x_t^1, \dots, x_t^N)$  and the observed one is  $\tilde{I} = f(\tilde{x}_t^1, \dots, \tilde{x}_t^N)$ .

We will select  $n < N$  data to be edited and we set  $d_j = 1$  when the  $j$ -th data is edited and  $d_j = 0$  otherwise. As a first approach for simplicity we assume that all data are correct after editing. If we put the  $d_j$ 's in a vector  $d = (d_1, \dots, d_N)$ , we can denote by  $\tilde{I}(d)$  the value of the statistic computed with the data edited according to  $d$ . Our aim is to decide which data will be edited in order to minimize certain measure of the error  $\tilde{I}(d) - I$ .

We denote by  $\mathcal{F}_t$  the  $\sigma$ -field that contains all the information available for editing, which comprises at least the rest of the data of the current period and from previous periods (but also other sources of information, random or deterministic, such as calendar information or even data from other surveys).

The selection problem can be stated as,

To find a the  $\mathcal{F}_t$ -measurable random variable  $d = (d_1, \dots, d_N)$  such as  $E(\tilde{I}(d) - I)^2$  is minimum subject to the constraint  $\sum d_i = n$ .

Let us assume that  $f$  is linear,  $f = \sum \omega_i x_t^i$ . Then,

**Proposition 1.** *The mean squared error  $E(\tilde{I} - I)^2$  attains a minimum when  $d_i = 1$  for the  $i$ 's with larger  $s_i$ , where,*

$$s_i = \omega_i^2 E[(\tilde{x}_t^i - x_t^i)^2 | \mathcal{F}_t] \quad (1)$$

23. The practical application of the proposition 1 requires to make some assumptions on the behavior of the data. We obtain the expectation in (1) by proposing a model for the true  $x_t^i$ . This model can be univariate or multivariate, since we can use data from other series to compute a prediction of  $x_t^i$ . Let us decompose  $\mathcal{F}_t$  as  $\mathcal{F}_t = \sigma(\tilde{x}_t^i, \mathcal{G}_t)$ . Thus, the prediction  $\hat{x}_t^i$  is a  $\mathcal{G}_t$ -measurable variable. We assume that,

- The prediction error  $\xi_t^i = x_t^i - \hat{x}_t^i = x_t^i - E[x_t^i | \mathcal{G}_t]$  is gaussian with variance  $\nu_i^2$ .
- The prediction error is independent of the observation error  $\varepsilon_t^i$  and both are jointly independent of  $\mathcal{G}_t$ .
- The distribution of the observation error is a mixture of a degenerate distribution (equal to zero) and a  $\sigma_i^2$ -variance gaussian, with probabilities  $1 - p$  and  $p$  respectively.
- The measurement error is independent of  $\mathcal{G}_t$ .

**Proposition 2.** *Under the assumptions above, it holds,*

$$E[(\tilde{x}_t^i - x_t^i)^2 | \mathcal{F}_t] = \left[ \sigma_i^2 + \frac{\sigma_i^4}{\sigma_i^2 + \nu_i^2} \left( \frac{u^2}{\sigma_i^2 + \nu_i^2} - 1 \right) \right] \zeta \quad (2)$$

where,

$$u = \hat{x}_t^i - \tilde{x}_t^i \quad (3)$$

$$\zeta = \frac{1}{1 + \frac{1-p}{p} \left( \frac{\nu^2}{\sigma^2 + \nu^2} \right)^{-1/2} \exp\left\{ -\frac{u^2 \sigma^2}{2\nu^2(\sigma^2 + \nu^2)} \right\}} \quad (4)$$

24. Sometimes, using transformed variables, such as the logarithm allows a more realistic modelling. Let us call  $y_t^i = \log(x_t^i)$ . Then, we can approximate  $E(\tilde{I} - I)^2$  by a first-order Taylor expansion as,

$$E(\tilde{I} - I)^2 \simeq \sum_i \omega_i^2 (\tilde{x}_t^i)^2 E[(y_t^i - \tilde{y}_t^i)^2] \quad (5)$$

Thus, we will select for editing the  $x_t^i$ 's corresponding to the larger  $s_i$ 's, where  $s_i$  equals now  $\omega_i^2 (\tilde{x}_t^i)^2 E[(y_t^i - \tilde{y}_t^i)^2 | \mathcal{F}_t]$ . This has several advantages,

- The assumption of gaussianity (and specially, the symmetry) of the observation error is more realistic for the logarithm.
- The variances  $\sigma_i^2$  and  $\nu_i^2$  are more uniform across  $i$ , so in case of lack of information about the behavior of the data, they can be assumed as constant.
- When the predictions are obtained through time-series models, they often use log-transformed variables

25. This approach allows to integrate different sources of information, such as time-series models or consistency checks using other variables of the same questionnaire.

## D. Results

26. In the INE HQ a database is maintained with all the versions of the questionnaire variables, as they are modified according to the macro/selective editing process. The selective editing method used is a simpler method described in (Arbués et al. 2005). We have used this information to simulate the editing process with the method described in this paper. The first version of the data is regarded as the *observed* data and the last version as the *true* data. We have used a year of data (from June

2005 to May 2006) to evaluate the performance of the method on the data after the first microediting. Then, we have assessed the performance on the raw data obtained from the web.

27. For this application, we have used the log version and the computation of the prediction  $\hat{y}$  consists on the selection of three very simple univariate models.

- $I(1)$ ,  $y_t = y_{t-1} + a_t$
- $I(1)_{12}$ ,  $y_t = y_{t-12} + a_t$
- $I(1) \times I(1)_{12}$ ,  $y_t = y_{t-1} + y_{t-12} - y_{t-13} + a_t$

Where  $a_t$  is gaussian white noise. To select the most adequate model of the three, we compute the residuals (i.e., the differenced series) and choose the model that produces less variance. This can be considered as a simplified automatic modelization system. The three models are selected with frequencies around 28%, 63% and 9%.

28. We have compared the performance of our score function with the following benchmark function, similar to the one defined, for example, in Hedlin (2003),

$$\delta_i^1 = \omega_i |x_t^i - \hat{x}_t^i| \quad (6)$$

where the prediction  $\hat{x}_t^i$  is the most recent value from previous periods of the survey.

Our score function  $\delta^3$  will be the  $s_i$  of proposition 1 with the conditional expectation of the logarithm version of proposition 2. The prediction  $\hat{y}$  is computed using the model selected as described above. In order to separate the gains obtained through the improved prediction from the ones obtained from the use of proposition 2, we have used an intermediate score function defined as in (6) but with the improved forecast. We call this intermediate function  $\delta^2$ .

Finally, we call  $\delta^0$  the *influence* function defined in Arbués et al. (2005) (the difference between the yearly rate of variation with and without using the data item; in case there is not available the  $t - 12$  value, the monthly rate is used instead).

The parameters  $p$  and  $\sigma$  have been estimated from the data as  $p = 0.001$  and  $\sigma = 2$  for Turnover and  $\sigma = 2.5$  for New Orders.

29. To compare the performance of different methods let us consider that we have the data arranged in decreasing order of score and define  $S(n) = (\sum_{j>n} \omega_i^2 (\hat{x}^i - x^i)^2)^{1/2}$ . This is an estimate of the Root Mean Square (RMS) of the remaining error when the data is checked up to  $n$ . The RMS accounts both for the standard deviation and for the bias. The ideal selection would produce a decreasing order of the terms in the sum. If we call  $S^*(n)$  the ideal curve, a simple metric of the goodness of the method can be  $\sum_n (S(n) - S^*(n))$ .

The result of computing the metric on the different methods for Turnover and for New Orders are in table 1.

TABLE 1. Measures of the score functions.

| Score function | Turnover | New Orders |
|----------------|----------|------------|
| $\delta^0$     | 496.2    | 648.4      |
| $\delta^1$     | 279.3    | 368.3      |
| $\delta^2$     | 209.3    | 229.6      |
| $\delta^3$     | 145.5    | 168.0      |

30. In tables 2 and 3 we present the fraction of squared error corrected and the RMS of the remaining error for different values of  $n$  and for Turnover and New Orders. In figure 2 we represent in logarithmic scale the fraction of remaining squared error vs. the fraction of data checked, and in

figure 3 we represent the root mean squared remaining error vs. the absolute number of data checked. With the exception of  $\delta^0$  for Turnover, figure 2 suggests that editing roughly 1% of the data removes around 90% of the squared error.

31. We have also evaluated the performance of the selective editing method to the data obtained through the web. The results are summarized in figure 4 and tables 4 and 5. The RMS values of figure 4 indicate that the web data have relatively more error (as expected, since they are unedited, while the other have passed the first micro-editing phase).

In order to have an operative measure of the performance of the method, we have estimated the number of data that we need to edit in order to reduce the RMS of the error under a desired threshold. Since the New Orders/Turnover indicators are rounded to one decimal place for publishing, 0.1 appears as a natural reference. By setting the threshold at 0.05, we would have a 95% confidence interval of half-width equal to 0.1 around the published value. For the whole sample data, we obtain this by editing around 2.5% of Turnover and 3.3% of New Orders.

The results of our experiment suggest that the selective editing allows to integrate the raw web data without great impact on the figures above and heavy editing work, since from tables 4 and 5, we can see that editing about 100 questionnaires (more accurately, 5.1% of the Turnover web data and 7.9% of New Orders), we have the error reduced to the order of 0.01.

#### IV. FINAL REMARKS

32. Nowadays, public statistical offices are under continued pressure from society, demanding more and more data to be produced at a lower cost, with a lower respondent burden, and especially with a shorter delay. Improving timeliness without any losses in accuracy is a major challenge for public statisticians today. Data editing should not only be linked to accuracy but also to others quality aspects, for example timeliness. Data editing is one of the most time-consuming statistical phases. Hence, re-engineering the data editing procedures is a need for improving timeliness.

Quality components are often considered as competing objectives. Nevertheless, new IT tools (such as Web questionnaires) and statistical methodologies (such as selective editing) offer the opportunity for re-engineering statistical production processes in order to improve some quality components simultaneously. According to our experience, the combination of Web questionnaires and selective editing strategy appears very promising.

#### References

- [1] Arbués, I., González-Dávila, M., González-Villa, M., Quesada, J., and Revilla, P. (2005) "EDR Impacts on Editing" UN/ECE Work session on statistical data editing (Ottawa, 2005)
- [2] Arbués, I., González-Dávila, M., González-Villa, M., Quesada, J. and Revilla, P. (2006) "Using a TQM approach to get high quality incoming data in the Spanish industrial surveys" European Conference on Quality in Survey Statistics.Q2006.
- [3] Branson, M. (2002) "Using XBRL for data reporting". Australian Bureau of Statistics. Unece/Eurostat Work Session on Electronic Data Reporting, Working Paper No 20. February 2002.
- [4] Gradjean, J.P. (2002) "Electronic data collection in Official Statistics in France". French National Institute of Statistics INSEE. UN/ECE/Eurostat. Work Session on Electronic Data Reporting. Working Paper No 7. February 2002.
- [5] Hedlin, D. (2003) "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics" *Journal of Official Statistics*, Vol. 19, No. 2.
- [6] Mayda, J. (2002) "Experiences with implementation of EDR into existing survey programs 2002". Statistics Canada. UN/ECE/Eurostat. Work Session on Electronic Data Reporting. Working Paper 23. February 2002.



- [7] Weir, P. (2005) "The Movement to Electronic Reporting and the Impact on Editing". 55th Session of the International Statistical Institute. Sydney. April, 2005
- [8] Weir P. (2005) "EDR and the impact on Editing - A summary and a case study" UN/ECE Work Session on Statistical Data Editing. Working Paper No 28. Ottawa. May 2005

## Appendix A. PROOFS

### *Proof of Proposition 1*

Let us consider a  $\mathcal{F}_t$ -measurable variable  $d^*$  under the assumptions. We can write,  $E(\tilde{I}(d^*) - I)^2$  as,

$$EE[(\tilde{I}(d^*) - I)^2|\mathcal{F}_t] = E \sum_i \omega_i^2 E[(\tilde{x}_t^i - x_t^i)^2|\mathcal{F}_t] \quad (7)$$

Since all data are correct after editing, and due to the  $\mathcal{F}_t$ -measurability of  $d^*$ , the expression above equals,

$$E \sum_i \omega_i^2 (1 - d_i^*) E[(\tilde{x}_t^i - x_t^i)^2|\mathcal{F}_t] \quad (8)$$

The sum in the expression above can be written as,

$$\mathcal{E} - \sum_i d_i^* s_i \quad (9)$$

It is easy to see that the subtracted term is greater when  $d_i^* = 1$  for the greater  $s_i$ 's.

### *Proof of Proposition 2*

In order to avoid heavy notation, we omit the superscript  $i$  and subscript  $t$ .

First, let us compute  $E[\varepsilon^2|u]$ . The error  $\varepsilon$  can be regarded as the product of a gaussian random variable  $\eta$  and a bernoulli  $e$  which equals 1 and 0 with probabilities  $p$  and  $1 - p$  respectively. Then,

$$E[\varepsilon^2|u] = E[e\eta^2|u] = E[E[e\eta^2|e, u]|u] \quad (10)$$

We can see that  $E[e\eta^2|e, u] = E[\eta^2|u]e$ . Since  $e$  is  $e$ -measurable,  $E[e\eta^2|e, u] = E[\eta^2|e, u]e = E[\eta^2|u, e = 1]e$  (where we denote by  $E[\eta^2|u, e = 1]$  the expectation  $\eta^2$  conditioned to  $u$  when there is error, i.e. when  $e = 1$ ). Then, we can write,

$$E[\varepsilon^2|u] = E[\eta^2|u, e = 1]E[e|u] \quad (11)$$

Now, in order to compute  $E[\eta^2|u, e = 1]$  we will use a well-known property of the gaussian distribution. If  $(x, y)$  is a gaussian-distributed random vector with zero mean and covariance matrix  $\Sigma = (\sigma_{ij})_{i, j = x, y}$ . Then, the conditional distribution  $f(y|x)$  is a gaussian with mean and variance,

$$E[y|x] = \frac{\sigma_{xy}}{\sigma_{xx}}x \quad (12)$$

$$V[y|x] = \sigma_{yy} - \frac{\sigma_{xy}^2}{\sigma_{xx}} \quad (13)$$

Then,

$$E[y^2|x] = \sigma_{yy} + \frac{\sigma_{xy}^2}{\sigma_{xx}} \left( \frac{x^2}{\sigma_{xx}} - 1 \right) \quad (14)$$

Now, we can apply the relation above to  $y = \eta$  and  $x = u = \eta + \xi$ . Since  $\eta$  and  $\xi$  are incorrelated,  $\sigma_{xx} = \sigma^2 + \nu^2$  and  $\sigma_{yy} = \sigma_{xy} = \sigma^2$ . Thus,

$$E[\eta^2|u, e = 1] = \sigma^2 + \frac{\sigma^4}{\sigma^2 + \nu^2} \left( \frac{u^2}{\sigma^2 + \nu^2} - 1 \right) \quad (15)$$

Let us now compute  $\zeta = E[e|u] = P[e = 1|u]$ . By Bayes's theorem, it is equal to  $p[e = 1]f(u|e)/f(u)$ , where  $f(u|e)$  is a zero-mean gaussian density function with variance  $\sigma^2 + \nu^2$  and  $f(u)$  is a mixture of two gaussians with variances  $s_1^2 = \sigma^2 + \nu^2$  and  $s_2^2 = \nu^2$  and probabilities  $p$  and  $1 - p$  respectively.

$$\zeta = p \frac{(2\pi s_1^2)^{-1/2} e^{-\frac{u^2}{2s_1^2}}}{p(2\pi s_1^2)^{-1/2} e^{-\frac{u^2}{2s_1^2}} + (1-p)(2\pi s_2^2)^{-1/2} e^{-\frac{u^2}{2s_2^2}}} \quad (16)$$

After simplifying it yields,

$$\zeta = \frac{1}{1 + \frac{1-p}{p} \left(\frac{\nu^2}{\sigma^2 + \nu^2}\right)^{-1/2} \exp\left\{-\frac{u^2 \sigma^2}{2\nu^2(\sigma^2 + \nu^2)}\right\}} \quad (17)$$

Finally, since  $\xi$  and  $\varepsilon$  are independent of  $\mathcal{G}_t$ , we can conclude that  $E[\varepsilon^2|u, \mathcal{G}_t] = E[\varepsilon^2|u]$ .

## Appendix B. FIGURES AND TABLES

TABLE 2. Remaining Error vs. Data checked.  $n$  = Number of data checked; % =  $100 \times n/N$ ; Rel SE = Squared Error Corrected / Total SE; RMSE = Root Mean Squared Error Remaining. Turnover (Whole sample).

| n     | %     | Rel SE | RMSE   |
|-------|-------|--------|--------|
| 1     | 0.01  | 0.2447 | 0.1778 |
| 10    | 0.07  | 0.5077 | 0.1435 |
| 25    | 0.18  | 0.6785 | 0.1160 |
| 50    | 0.36  | 0.6942 | 0.1131 |
| 100   | 0.71  | 0.7628 | 0.0996 |
| 200   | 1.43  | 0.8807 | 0.0706 |
| 300   | 2.14  | 0.9367 | 0.0514 |
| 400   | 2.86  | 0.9407 | 0.0498 |
| 500   | 3.57  | 0.9505 | 0.0455 |
| 1000  | 7.14  | 0.9634 | 0.0391 |
| 2000  | 14.29 | 0.9889 | 0.0216 |
| 3000  | 21.43 | 0.9900 | 0.0205 |
| 4000  | 28.57 | 0.9993 | 0.0053 |
| 5000  | 35.71 | 0.9997 | 0.0034 |
| 10000 | 71.43 | 1.0000 | 0.0002 |

FIGURE 2. Relative effectiveness of the method. The x-axis represents the fraction of data checked and the y-axis represents the SE remaining. Both axis are in logarithmic scale.

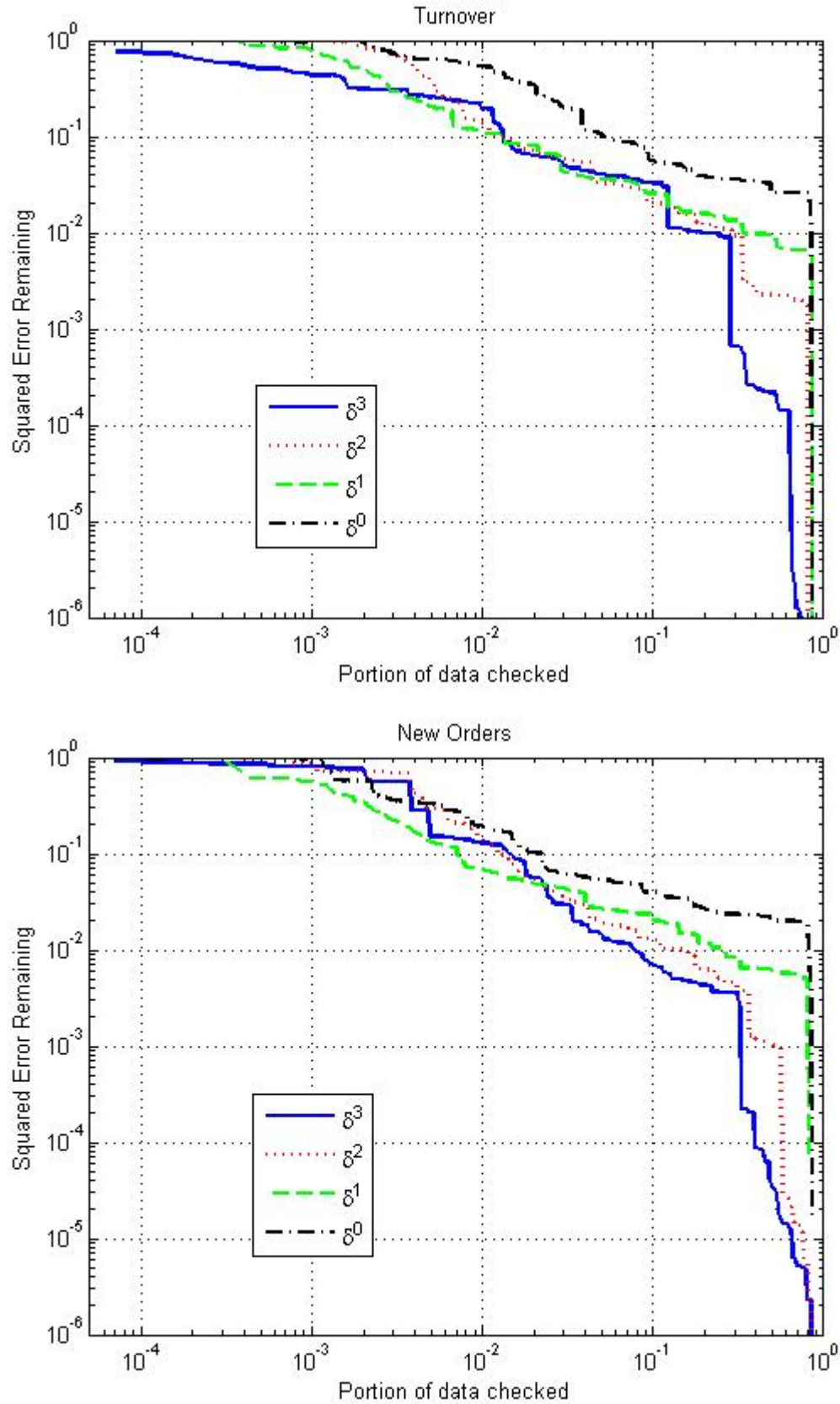


TABLE 3. Remaining Error vs. Data checked.  $n$  = Number of data checked; % =  $100 \times n/N$ ; Rel SE = Squared Error Corrected / Total SE; RMSE = Root Mean Squared Error Remaining. New Orders (Whole sample).

| n     | %     | Rel MSE | RMSE   |
|-------|-------|---------|--------|
| 1     | 0.01  | 0.0902  | 0.3203 |
| 10    | 0.07  | 0.1942  | 0.3015 |
| 25    | 0.18  | 0.2372  | 0.2933 |
| 50    | 0.36  | 0.4419  | 0.2509 |
| 100   | 0.71  | 0.8575  | 0.1267 |
| 200   | 1.43  | 0.8998  | 0.1063 |
| 300   | 2.14  | 0.9434  | 0.0799 |
| 400   | 2.86  | 0.9699  | 0.0582 |
| 500   | 3.57  | 0.9801  | 0.0473 |
| 1000  | 7.14  | 0.9883  | 0.0363 |
| 2000  | 14.29 | 0.9951  | 0.0236 |
| 3000  | 21.43 | 0.9957  | 0.0219 |
| 4000  | 28.57 | 0.9964  | 0.0201 |
| 5000  | 35.71 | 0.9998  | 0.0049 |
| 10000 | 71.43 | 1.0000  | 0.0008 |

TABLE 4. Remaining Error vs. Data checked.  $n$  = Number of data checked; % =  $100 \times n/N$ ; Rel SE = Squared Error Corrected / Total SE; RMSE = Root Mean Squared Error Remaining. Turnover (Raw web data).

| n    | %     | Rel MSE | RMSE   |
|------|-------|---------|--------|
| 1    | 0.07  | 0.0512  | 0.1626 |
| 10   | 0.71  | 0.9304  | 0.0440 |
| 25   | 1.79  | 0.9556  | 0.0352 |
| 50   | 3.57  | 0.9868  | 0.0192 |
| 100  | 7.14  | 0.9996  | 0.0034 |
| 200  | 14.29 | 0.9998  | 0.0026 |
| 300  | 21.43 | 0.9999  | 0.0015 |
| 400  | 28.57 | 0.9999  | 0.0015 |
| 500  | 35.71 | 0.9999  | 0.0012 |
| 1000 | 71.43 | 1.0000  | 0.0000 |

TABLE 5. Remaining Error vs. Data checked.  $n$  = Number of data checked; % =  $100 \times n/N$ ; Rel SE = Squared Error Corrected / Total SE; RMSE = Root Mean Squared Error Remaining. New Orders (Raw web data).

| n    | %     | Rel MSE | RMSE   |
|------|-------|---------|--------|
| 1    | 0.07  | 0.0999  | 0.0973 |
| 10   | 0.71  | 0.5673  | 0.0675 |
| 25   | 1.79  | 0.5814  | 0.0664 |
| 50   | 3.57  | 0.8847  | 0.0348 |
| 100  | 7.14  | 0.9903  | 0.0101 |
| 200  | 14.29 | 0.9948  | 0.0074 |
| 300  | 21.43 | 0.9990  | 0.0032 |
| 400  | 28.57 | 0.9992  | 0.0029 |
| 500  | 35.71 | 0.9995  | 0.0023 |
| 1000 | 71.43 | 1.0000  | 0.0000 |

FIGURE 3. Root Mean Square of the remaining errors vs. data checked.

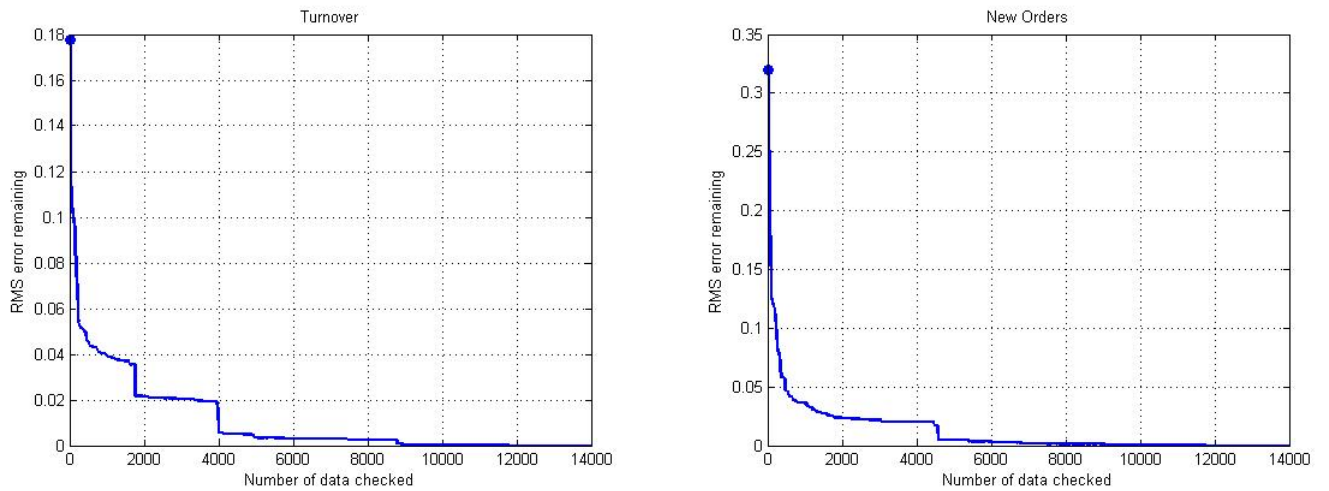


FIGURE 4. Root Mean Square of the remaining errors vs. data checked (web).

