WP. 21
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

Evolving Data Processing in the Statistics Centre – Abu Dhabi

Prepared by Dragica Sarich and Maitha Al Junaibi,
Statistics Centre – Abu Dhabi, United Arab Emirates

Abstract: Official statistical agencies commonly use statistical techniques to improve the quality of data, such as data editing. These statistical techniques provide several benefits for official statistical agencies, including allowing for 'complete' data to be obtained (i.e., no missing responses for items), and enabling further statistical analyses to be performed. Abu Dhabi's official statistical office - Statistics Centre – Abu Dhabi (SCAD) - has made ground-breaking progress for the Emirate of Abu Dhabi by conducting its Economic Surveys using mixed-mode data collection and automated data error detection for the first time, and performing some new methods of statistical techniques to this data to enhance data quality. This paper discusses the advantages and disadvantages of performing data editing, SCAD's experiences with applying automated data error detection as part of data processing, and strategies implemented to overcome issues encountered.

I. Introduction

1. Data editing is an important process of data processing undertaken by international official statistical agencies. Data editing is performed to address errors or inconsistencies introduced into survey data by respondents or through, field or data entry errors (De Waal, Pannekoek, & Scholtus, 2011; International Household Survey Network, 2009). Data editing is a three-fold process involving: 1) reviewing data collected from field, 2) identifying data entry errors or inconsistent data according to a set of rules, and 3) correcting the errors or inconsistencies (De Waal, Pannekoek, & Scholtus, 2011; International Household Survey Network, 2009). Data editing can be an intensive and time-consuming task, if done manually, hard-coded to unit-specific records, or requiring units to amend their own responses. For official statistical agencies, this adds to production costs, reduces staff availability and extends timelines. For units, it can contribute to or even increase respondent burden. Given this, of key interest to official statistical agencies and researchers (such as Pannekoek, 2009) is how to evolve the process of data editing and, in turn, experience added benefits for all parties concerned. Benefits

include: data quality can improved with relative ease, staff availability can be increased and devoted to other statistical duties/ projects, production costs for deliverables can be reduced, and decreases respondent burden.

2. The Statistics Centre – Abu Dhabi (SCAD) is a relatively new official statistical agency for the Emirate of Abu Dhabi. This statistical agency recently commenced its program of Economic Surveys (including the Annual Economic Survey (AES) (2010-11)[1], Foreign Investment Survey (FI), and Yearly Environmental Survey (YES)), using mixed-mode data collection involving the use of web, telephone, email, mail and facsimile to gather data required from establishments for the purpose of deriving economic statistical indicators, such as Gross Domestic Product (GDP). To evolve the data editing required being undertaken for these surveys and experience benefits such as those mentioned above, SCAD developed and introduced to the Emirate an automated data error detection system for establishment surveys. In its efforts to further improve the quality of data being collected through its economic surveys, SCAD also applied techniques to assist in the statistical treatment of missing and anomalous data. This paper will discuss the advantages and disadvantages of performing data editing and statistical techniques on erroneous, anomalous and/or missing data (with a focus on economic data), SCAD's experiences with the application of the automated system and statistical techniques, and strategies implemented to overcome issues encountered and to improve data quality.

II. Data Editing

A. Advantages and disadvantages of data editing performed manually or hard-coded to unit-specific records

3. Data editing performed manually involves a field officer being provided with data in hard copy questionnaires and a list of editing rules which they are to use to make comparisons against as well as across questions as a method for identifying data errors or inconsistencies. The field officer then makes the correction on the hard copy questionnaire (for record keeping purposes), then keys the correction into the data entry system (United Nations, 2001). As an alternative means, data editing performed via hard-coding in statistical programs for unit-specific records involves writing a program/s to identify and correct an error for a particular unit/s. For example, in SAS, this may be:

```
proc print data= AES_version1;
        where q001<500;
        var est_id q001;
        title 'Check of establishments whose response to question 1 is less than AED500';
run;
```

[1] n = 4,000, n = 2,000 and n = 200, respectively.

```
data data.AES_version2;
set AES_version1;
        if id = '100a' and q001=50 then do; *error identified for establishment 100a;
                q001=500; *error corrected for establishment 100a;
        end;
run;
```

4. Data editing performed either manually or hard-coded in statistical programs for unit-specific records presents several advantages and disadvantages. The advantages include:

- inconsistencies in data can be identified and resolved, which can for some individual cases improve the quality of the data; and
- the overall quality of the data can improve.

5. The disadvantages include:

- labour-intensive;
- reduces staff resource availability for other aspects of the project;
- time-consuming for the person making the edit (be it the unit, the data reviewer, or the survey analyst);
- increases respondent burden (if the respondent is the one making the edit or providing the correction to the data);
- provides significant scope for the introduction of further human errors;
- can decrease the quality of the data for individual cases; and
- increases production costs for producing deliverables (Al-Hamad, Silva & Brown, 2007; United Nations, 2001).

Thus, it can be seen that data editing performed manually or hard-coded to unit-specific records can place great demands on organisations, including official statistical organisations.

B. Advantages and disadvantages of automated data error detection and editing

6. Automated data error detection and editing presents several advantages to official statistical agencies, including:

- computers are not affected by certain factors that can affect staff productivity such as exhaustion and personal problems that might interfere with maintaining data quality of consistency;
- computers are faster at detecting errors and, consequently, speed up data processing as well as reduce production costs of deliverables;
- complex or detailed consistency checks can be established and developed with subject matter experts; and
- automated data error detection can look at many characteristics simultaneously.

Automated data error detection and editing also minimizes respondent burden for units or establishments, as they are often not required to review and resubmit their own data.

7. While these advantages are beneficial and value-adding to statistical organizations and to respondents, statistical organizations should remain cognizant that automated data error detection

and editing requires quality control (United Nations, 2001) from two aspects: 1) over the quality of incoming data and 2) in terms of the code used which forms the automated system. The latter case is very important to ensure, as a small error in the code can reduce the quality of the functionality of the system. For example, the system may not identify and flag cases with errors or may flag cases that are not errors. If statistical organizations can devote resources to this element of quality control, then automated error detection and editing can provide the benefits identified above as well as evolve the business of the statistical organization in terms of data editing and processing.

III. Automated Data Error Detection – Application by SCAD

A. SCAD and its program of economic surveys

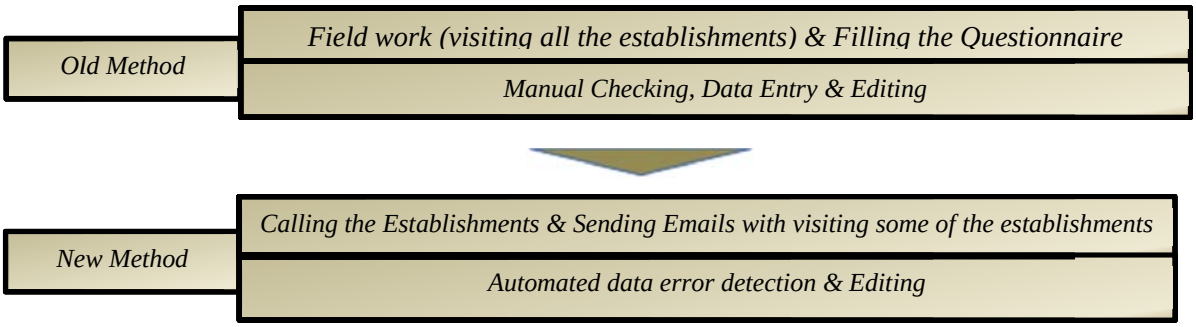
8. SCAD is the official statistical agency for the Emirate of Abu Dhabi, United Arab Emirates (UAE) (Statistics Centre – Abu Dhabi, 2012a). This agency commenced operation in April 2009 (Statistics Centre – Abu Dhabi, 2012a, 2012b). Part of SCAD's roles and responsibilities as an official statistical agency include: preparing statistical plans and programs that serve the Emirate's development programs; conducting a variety of statistical surveys across the Emirate, including social, demographic, economic, environmental and cultural surveys; and collecting, classifying, storing, analysing and disseminating official statistics including releasing results from surveys undertaken (Statistics Centre – Abu Dhabi, 2012a). Outputs produced by SCAD serve as a key source of information to decision and policy makers, researchers, the business community, the public, and the media within the UAE (Statistics Centre – Abu Dhabi, 2012a).

9. Since commencing operation, SCAD has conducted a program of Economic Surveys incorporating and including the Annual Economic Survey (AES), Foreign Investment Survey (FI), and the Yearly Environmental Survey (YES). These surveys are key data sources for the compilation of economic, foreign investment, and environmental statistics essential to the Government of the Emirate of Abu Dhabi (Statistics Centre – Abu Dhabi, 2012a). These surveys collect data from establishments spread across three regions of the Emirate (Abu Dhabi, Al Ain and Al Gharbia) on an annual basis, and aim to measure:

- the structure and performance of sectors within the Emirate of Abu Dhabi's economy;
- the volume, flow, source and role of foreign investments in the Emirate; and
- environmental and health and safety issues.

10. Prior to 2011, SCAD collected data for the program of Economic Surveys and other surveys using face-to-face interviewing. With the expectation of an intensive program of Economic Surveys for 2011, SCAD sought to expand its options and capabilities in data collection through the use of mixed mode data collection and, for data editing and processing, automated error detection. The figure below shows the difference between the old and new methods of data collection and data editing.

Figure 1. Old and new methods of data collection and editing in SCAD.



B. The design and application of mixed mode data collection and automated data error detection systems by SCAD to its program of economic surveys

11. Mixed mode data collection for each survey in SCAD's program of economic surveys involved the development and design of paper-based questionnaires through liaison and consultation with internal subject matter experts of the surveys. These questionnaires were then designed in Sensus[2], a system for the creation and management of online questionnaires. The samples for each survey were uploaded into WinCati[3]. The online questionnaires were deployed to establishments via email and establishments were informed and provided with the option to complete the surveys via the web or complete the hardcopy questionnaire and either mail or fax the completed questionnaire to SCAD. Using WinCati, SCAD contacted each establishment in the sample for introductions, reminders, and answering establishments' queries.

12. Further to this, for each of these surveys using SAS Enterprise Guide and R[4], SCAD developed automated data error detection systems to check data provided by establishments and to identify and flag erroneous data featured in the unit record file. Like the survey questionnaires, these systems were designed and developed through liaison and consultation with internal subject matter experts. Since the AES, FI and YES are repeat surveys the internal subject matter experts already had experience and insight into expected responses to questions featured in the survey questionnaires and into the quality of the data. From this, the subject matter expects developed a set of validation rules for each survey which acted as a guideline for checking the data, and identifying and flagging erroneous or anomalous data, as well as for making edits.

13. Automated data error detection systems were then developed in SAS and R. A team of SCAD's survey analysts translated the validation rules provided by the internal subject matter

---

[2] Sawtooth Technologies, http://www.sawtooth.com/index.php/software/products/sensus-web/
[3] Sawtooth Technologies, http://www.sawtooth.com/index.php/software/products/wincati/
[4] R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

experts into these statistical packages' programming languages. An example of a validation rule developed for the AES 2011 along with SAS code is as follows:

- *License number must be less than 8 digits*

```
proc print data=status4v1RR;
     var EST_ID respnum Q03101 Q0201;
    where (Q03101 not in (1,2,3) and Q0201 > 9999999) or Q0201 in (.,0);
    title "Validation:  License number more than 7 numbers  [or is '.' or '0']";
run;

    *set flag;
data status4v2;
set status4v1RR;
      if (Q03101 not in (1,2,3) and Q0201 > 9999999) or Q0201 in (.,0) then
      FIELDCHECK_Q0201_CRITICAL=1;
run;
```

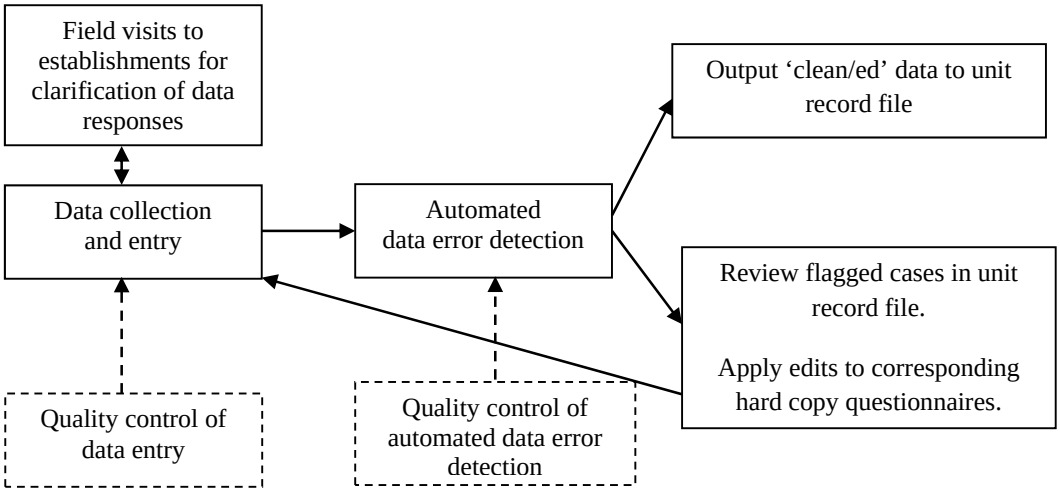
14. SCAD's team of survey analysts also programmed and created identifier variables or 'flags' in the data file to identify those records that did not meet each validation rule and/or did not pass what was deemed as critical or tolerable validation rules by the experts. These automated data error detection systems also were built with the functions of performing outlier detection, managing the coding of open-ended responses, producing reports and logs of outcomes for each establishment based on the validation checks, and preparing a unit record file. In line with United Nation's (2001) awareness and encouragement about quality control with automated data error detection and editing, two teams of statistical experts within SCAD performed quality assurance on the automated data error detection system for each survey.

15. These automated data error detection systems were then incorporated into the collection and processing phases for each of the three surveys. Specifically, following deployment of the surveys into field, data was retrieved from establishments via mixed mode data collection. For data received via fax or mail, data entry was performed internally using the online questionnaires. This allowed data being collected via the several means to be stored in one central location – Sensus. Data was then passed through the automated data error detection systems from which establishments' records flagged with potentially erroneous or anomalous data were reviewed and manually edited (using mail-merged paper based questionnaires) by SCAD's team of subject matter experts. These establishments' hardcopy questionnaires were then returned to field officers for rekeying the edit/s into Sensus. As a result of this review, some establishments were re-contacted by field officers face-to-face about data responses provided and for the provision of data correction. Field officers then rekeyed the edit/s into Sensus.

16. As the automated data error detection system was run on a daily basis, those establishments' data that had previously been flagged, reviewed, and edited would now no longer be flagged by the automated data error detection system the subsequent time/s around and pass through the system and sit in the unit record file as a 'cleaned' record. As a cross-check of

the newly implemented system for validation checks, two teams were set up within SCAD focusing on checking the quality of data entry and checking that the automated data error detection system produced the same outcomes as manual error detection according to the devised set of validation rules provided by the subject matter experts. In future, it is envisaged that such teams will be small and will only handle special cases in the data. The diagram below presents the implemented data cycle for the AES, FI and YES.

Figure 2. Implemented data cycle for AES, FI and YES



C. Issues encountered and strategies implemented by SCAD

17. A key issue was initially encountered in the development of the automated data error detection system – a large number of validation rules were set which could potentially increase respondent burden, increase respondent fatigue and attrition, decrease the quality of the data through an increase in non-response and in respondents' reluctance to assist with data corrections, and jeopardize response rates. To overcome this issue, with much liaison and consultation within SCAD across specialist teams, the strategic approach was to: 1) review the set validation rules, 2) remove those deemed of a low weight to the subject matter experts, 3) prioritize the remaining validation rules based on those perceived to be critical to the subject matter experts and those considered to be tolerable if failed, 4) where necessary, create and incorporate new or consolidated rules into the automated data error detection systems. This strategy proved to be a success. A revised and smaller set of validation rules was produced and incorporated into the automated data error detection systems. Also, in the case of the AES 2010-11, data inflow obtained by field from establishments has been voluminous (approaching approximately 90% consent rate to-date) being achieved via the mixed mode data collection. Moreover, markedly fewer establishments were flagged for requiring reviewing and editing based on the validation rules than previously was the case. This in turn reduced respondent burden for many establishments and thus minimized respondent attrition in the surveys, and freed up staff resources – allowing attention to be devoted to the application of other statistical techniques for improving the quality of the data.

IV. Conclusion

18. Data editing is an important phase of data processing. As discussed in this paper, the development of an automated system for elements of data editing, such as error detection in the case of SCAD, can provide many benefits to both statistical agencies and respondents.

19. SCAD is currently investigating several methods for the statistical treatment of missing or anomalous data in establishment surveys – even with the best design and methods of collection, such issues are always present in survey data. The aim of this process is to develop a system that is robust, simple to implement, and both comprehensible and useful to the subject matter experts tasked with producing economic indicators.

20. SCAD encourages other statistical agencies to take on board SCAD's experiences with evolving data processing and consider the view of incorporating automation into phases of survey projects.

V. References

Al-Hamad, A., Silva, P., & Brown, G. (2007). *Improving the efficiency of data editing and imputation for a large-scale British annual business survey.* Retrieved from: http://www.fcsm.gov/07papers/Al-Hamad.III-B.pdf

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Statistical Data Editing and Imputation.* Wiley: New Jersey.

International Household Survey Network. (2009). *Data Editing.* Retrieved from: http://www.surveynetwork.org/home/index.php?q=tools/editing/principles

Pannekoek, J. (2009). *Research on edit and imputation methodology: The throughput programme*. Statistics Netherlands, The Hague.

Statistics Centre – Abu Dhabi. (2012a). *Statistics Centre – Abu Dhabi: An overview*. Retrieved from: www.scad.ae/en/aboutus/Pages/Overview.aspx

Statistics Centre – Abu Dhabi. (2012b). *Statistics Centre – Abu Dhabi: Director-General Message*. Retrieved from: intranet/en/Pages/default.aspx

United Nations. (2001). *Handbook on Population and Housing Census Editing*. Retrieved from: http://www.un.org/Depts/unsd/