

REAL-TIME DYNAMIC EDITING – A PRELIMINARY STUDY OF SOME METHODOLOGICAL ISSUES

Sanping Chen and Paul Hunsberger¹

ABSTRACT

Real-time editing within the data collection process allows early detection and follow-up of problematic records, improving the operational efficiency and timeliness. The challenge is the lack of the full knowledge of the current data, as required by almost all data editing and/or outlier detection techniques. Real-time editing tries to use estimates of individual current records, which are gradually replaced by the true values in the collection process. This can be viewed as some sort of sequential decision process.

Technical issues regarding real-time editing are: The uncertainty level and its relationship with the collection sequence, dynamic boundaries based on collection time, retrospective decisions and their impact, and some group-sequential methods whenever editing is conducted at the end of a day or a week.

KEY WORDS: Real-time editing; Dynamic editing; Selective editing.

RÉSUMÉ

La vérification en temps réel dans le processus de collecte permet une détection et un suivi immédiats des enregistrements problématiques, ce qui améliore l'efficacité et les délais opérationnels. Le défi est le manque de connaissances des données actuelles, comme le requiert presque toute technique de vérification et de détection de données aberrantes. La vérification en temps réel tente d'utiliser des estimations des enregistrements actuels, qui sont graduellement remplacées par de vraies valeurs dans le processus de collecte. Ceci peut être vu comme un processus séquentiel de décision.

Les problèmes techniques liés à la vérification en temps réel sont : le niveau d'incertitude et sa relation avec la séquence de collecte, les bornes dynamiques basées sur la période de collecte, les décisions rétrospectives et leurs impacts, et quelques méthodes de traitement par groupes lorsque la vérification n'est faite qu'à la fin d'une journée ou d'une semaine.

MOTS CLÉS: Vérification en temps réel; Vérification dynamique; Vérification selective.

1. INTRODUCTION

Data editing is often the most costly and time-consuming step in the business survey process. For example, the United States Federal Committee on Statistical Methodology reported in 1990 that in the federal statistical agencies, the median editing costs for economic surveys were 40% of the total survey cost. According to Latouche and Berthelot (1992), recontact and follow-up of sample units are the most expensive tasks in the collection and capture process.

Several methods have been proposed to reduce the cost of data editing, of which selective editing is

perhaps the most important. In a nutshell, selective editing tries to minimize the follow-up task by concentrating the recontact effort on sample units that may have a significant effect on the major survey estimates. In other words, it is an effort to optimize the amount of resources used in data editing. Technically, selective editing decisions are based on carefully developed score functions using historical and current data.

By its rationale of examining the effect on survey estimates, it is naturally implied that the score function for selective editing always involves sample aggregates. This fact is noted by Latouche and

¹ Authors are from Business Survey Methods Division, R.H. Coats Building 3rd Floor, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada. E-mail address of the first author: Chensan@statcan.ca.

Berthelot. In order not to delay the survey process, they have suggested that these parameters be based on data from prior survey cycles. However, as we shall demonstrate through numerical examples, such an approach reduces the relevance of the purpose of the score function: to measure the effect on *current* survey estimates.

If we go beyond selective editing, which requires the availability of historical data, we shall see that the knowledge of full current-cycle data is generally required for data editing. We can cite the oldest and simplest outlier detection method, in which an outlier is identified if a record is more than, say 2.5 or 3 standard deviations away from the mean. The score function here for the i -th record is the simple z-score based on the sample mean \bar{x} and sample standard deviation s :

$$Z_i = \frac{x_i - \bar{x}}{s}$$

In fact, the Generalized Edit and Imputation System (GEIS) in use at Statistics Canada adopts a nonparametric variation of the above, in which the sample median and sample interquartile range replace the sample mean and standard deviation respectively, as the “current method” of outlier detection.

Real-time Dynamic editing is a new editing method for handling the general editing problem in which the score function requires the knowledge of the full current-cycle data. In applications involving historical data, real-time editing can be viewed as a further extension of selective editing. But in contrast to selective editing which tries to minimize the amount of resources in data editing, real-time editing represents an effort to better and more efficiently allocate resources, by allowing early detection and follow-up of problematic records within the data collection process, thus improving the operational efficiency and timeliness.

In a nutshell, real-time editing uses a dynamic score function for outlier detection. This score function is calculated and updated every time a new record is collected. In most applications extended from a selective editing method, this score function is simply the old score function calculated using a dynamic pseudo-dataset. This pseudo-dataset consists of current-cycle records collected so far, plus the “proxies” or estimates of yet-to-be-collected individual current-cycle records, which are gradually replaced by the true records in the collection process. More specifically, after the collection of the i th record, the pseudo-dataset looks like

$$(x_1^{(t)}, x_2^{(t)}, \dots, x_i^{(t)}, \hat{x}_{i+1}^{(t)}, \hat{x}_{i+2}^{(t)}, \dots, \hat{x}_n^{(t)}),$$

where t represents the current cycle, and \hat{x} 's are proxies.

While most of our examples follow the dynamic pseudo-dataset method, it should be noted that this is not the only approach for real-time editing. For example, in the simplest current method mentioned above, the dynamic score can be simply based on the up-to-date partial sample mean and sample standard deviation. Also, it should be pointed out that our examples here are applied to only one variable at a time, whereas extension to several variables concurrently would be a definite possibility for expansion of the process.

Real-time Dynamic editing bears a strong resemblance to the sequential method in statistical hypothesis testing, in that first it is a sequential decision process and secondly the decision is based on information collected up to the time point. This resemblance naturally leads to non-constant control boundaries and “group sequential” methods (when data is processed in batch, say daily or weekly). Nonetheless, there is a key difference between sequential hypothesis testing and Dynamic editing, in that the latter allows retrospective decisions. In other words, in Dynamic editing, once a new record is added, the score function for all previous records which have not yet been classified as an outlier is also updated. Therefore, one can reconsider decisions made on the previous records, based on the updated score function.

The benefits of real-time editing are apparently manifold. The most important benefit is the time savings accorded by early follow-up. In addition, more timely follow-up is more likely to produce more accurate data. Real-time recontacting may also help lower survey costs directly. It also provides a general and better approach to deal with sample rotations, births and deaths that are difficult with selective editing based on historical data. In general, real-time editing improves the overall timeline and cost-efficiency of the survey process.

2. AN EXAMPLE BASED ON THE LATOUCHE-BERTHELOT SELECTIVE EDITING METHOD

Here we use the score function method Latouche and Berthelot (1992) have proposed for selective editing to demonstrate Dynamic editing. Due to space considerations, we shall use only the “difference score”. Also it is worth noting that for simplicity, all potential survey weights w_i have been set=1 for this

example and thus do not affect the results. For record i , this difference score is essentially

$$S_i = \frac{|x_i^{(t)} - x_i^{(t-1)}|}{Y}$$

where Y is the aggregate. Evidently this aggregate should be of the current cycle (t) to most accurately measure the impact on the statistics being produced. We call such a score the "current method" score:

$$S_{i(C)} = \frac{|x_i^{(t)} - x_i^{(t-1)}|}{Y^{(t)}}$$

However, such a score requires the full knowledge of current data and cannot be produced "real time" during data collection. To overcome this difficulty, Latouche and Berthelot propose to use the aggregate from the previous cycle ($t-1$), generating what we call the Latouche-Berthelot method score, useable in "real time":

$$S_{i(LB)} = \frac{|x_i^{(t)} - x_i^{(t-1)}|}{Y^{(t-1)}}$$

For Dynamic editing, we propose to use a dynamic score based on the pseudo-dataset:

$$S_{i(D)} = \frac{|x_i^{(t)} - x_i^{(t-1)}|}{\hat{Y}_i}$$

where \hat{Y}_i is calculated dynamically for each i .

We applied all three methods to Statistics Canada's Annual Wholesale and Retail Trade Survey data (variable no. 61), using simple substitution by the previous cycle record as the proxy to establish the pseudo-dataset for Dynamic editing. Because the dynamic score can vary depending on the collection sequence of the records, we took the first 1000 records

and randomly scrambled them 25 times to measure the average performance of the dynamic method. The results are shown graphically in the attached chart. The current method was used as the basis of comparison, considering it to be the most relevant method for detecting outliers since it has 100% of the collected data at its disposal. The result is therefore a scatter plot of the difference between each score value and the current score by the record sequence number. For comparison between the Dynamic method versus Latouche-Berthelot method, best-fit lines representing the average difference of the 25 runs for each method were plotted in the following chart.

The results indicate that Dynamic editing provides score values that approximate the current method scores more and more closely as the data collection cycle progresses. This allows for an increasingly accurate detection of outliers and represents a significant improvement on existing real-time editing techniques.

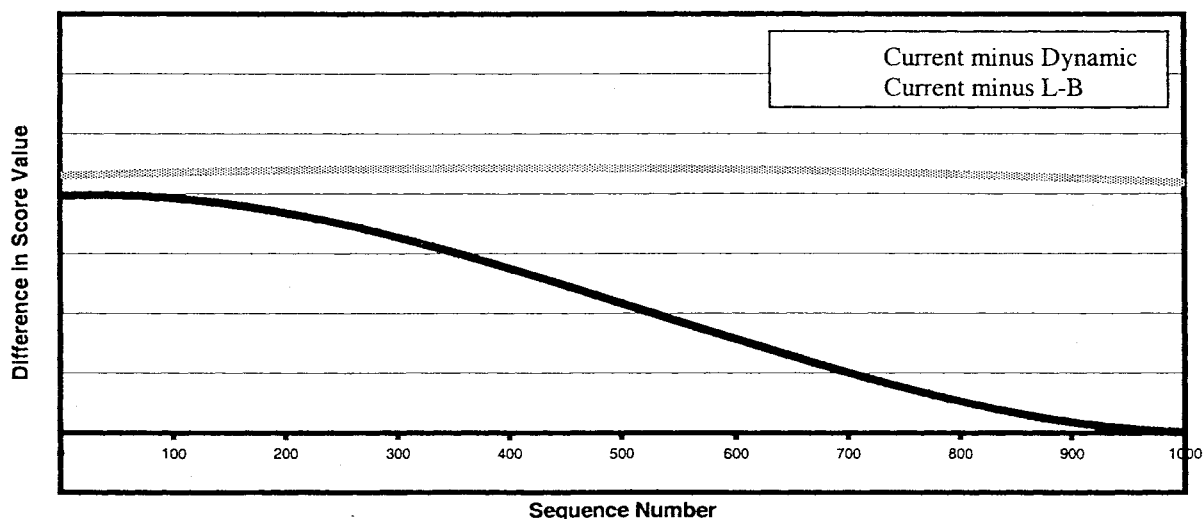
3. THE ISSUE OF PROXY RECORDS

As mentioned earlier, the easiest method for real-time editing is often to utilize the existing score function but to calculate it on a dynamic pseudo-dataset, in which all the yet-to-be-collected records are represented by proxies. Also as shown in our previous example, the simplest or most naïve proxy is just to substitute the record from the previous cycle:

$$\hat{x}_i^{(t)} = x_i^{(t-1)}$$

Because most business surveys are conducted in an ever-changing market environment, one can certainly develop more sophisticated proxy models to reflect

Comparison of Editing Methods
(Current, Latouche-Berthelot, and Dynamic)



this critical nature of business surveys. For instance, one can introduce a time trend term in the proxy:

$$\hat{x}_i^{(t)} = x_i^{(t-1)} + \mu.$$

At each step i , the time trend is estimated dynamically by

$$\hat{\mu} = \frac{1}{i} \sum_{j=1}^i (x_j^{(t)} - x_j^{(t-1)}).$$

(We have also found that a non-parametric alternative in which the trend is estimated by the median difference leads to better results when the data distribution is highly skewed).

The time-trend model is a special case of a general linear model using historical data:

$$\hat{x}_i^{(t)} = \alpha + \beta_1 x_i^{(t-1)} + \beta_2 x_i^{(t-2)} + \dots,$$

whose parameters can be estimated dynamically as in the time-trend model.

In general, a proxy model can involve not only historical data but also auxiliary variables:

$$\hat{x}_i^{(t)} = f(x_i^{(t-1)}, x_i^{(t-2)}, \dots, z_i, \dots).$$

The parameters in the function are to be estimated and updated dynamically based on partial current data and full historical data. The functional form may not necessarily be linear, as businesses seldom change linearly. Multiplicative and exponential models may be more appropriate in certain cases. The inclusion of auxiliary variables may also be of help in unique real

time situations where the expected set of historical data is not completely available, such as in processing sample rotations, births and deaths. These situations are difficult to handle using existing selective editing methods. Another possibility is to develop a dynamic score function by combining current and historical methods for outlier detection.

4. OPTIMAL CONTROL BOUNDARY

A key motivation for real-time editing is early outlier detection and recontact. Naturally, the earlier an outlier is detected, the bigger the savings in survey process time. However, the cumulative aspect of Dynamic editing means that earlier decisions will be

based on less information, and hence will be subject to a greater likelihood of incorrect decisions. This is further complicated by retrospective decisions, in which the score function for every previous record which has not yet been classified as an outlier is also updated each time a new record is added. All this leads to the question of an optimal control boundary to best balance the time savings against incorrect decisions, given the existence of retrospective decisions.

Here we present a formulation of this optimization problem, which is naturally one of many possibilities to define and construct the problem.

Let us assume that records are processed sequentially at step $i=1,2,\dots,n$. At each step i , k_i new records are processed (general batch-process problem). Without loss of generality, assume a one-sided problem in which an outlier is defined as some score function S exceeding a fixed boundary B : $S > B$.

All possible random permutations of the record entry sequences generate the probability space for our problem. Given the dynamic score S' based on, say, some chosen proxy model, the optimization problem is how to choose the control boundaries B'_1, B'_2, \dots, B'_n , at each step such that some total benefit is maximized.

Because of the extensive variety of possible editing methods and corresponding score functions, we have found it better to parametrize the problem using not the control boundaries directly, but the proportion or probability of resulting outlier calls in the probability space defined above:

$$p_i = \Pr(S'_i > B'_i).$$

A very important related quantity is the probability of making an incorrect outlier call:

$$h_i = \Pr(S_i \leq B \mid S'_i > B'_i),$$

which is determined jointly by the choice of score function, the proxy model, if any, and the dynamic control boundary B'_i . [Where S_i is the "current" score at step i , and B is the fixed "current" control boundary].

Let us assume that the penalty for an incorrect outlier call (i.e., a wasted follow-up) is a constant C . We also assume that the reward for a correct outlier call at step i is proportional to the time saved, namely $R(n-i)$. It can be shown that the expected total benefit of the whole process is

$$\sum_{i=1}^n E_{i-1}(K_i) p_i \{(1-h_i)R(n-i) - h_i C\},$$

where $E_{i-1}(K_i)$ is the expected number of records processed at step i (new records collected at step i plus all previous records that have not been identified as outliers yet), which can be shown to be

$$E_{i-1}(K_i) = k_i + k_{i-1}(1-p_{i-1}) + k_{i-2}(1-p_{i-1})(1-p_{i-2}) \\ + \dots + k_1(1-p_{i-1})(1-p_{i-2}) \dots (1-p_1).$$

An optimal boundary is a set of $(B'_1, B'_2, \dots, B'_n)$, or equivalently a set of (p_1, p_2, \dots, p_n) , that maximizes the expected total benefit.

Evidently, the optimal boundary is determined by many factors including the choice of score function and of the proxy model, and thus is not an easy problem. Analytically, the functional relationship between p_i and h_i is of critical importance in solving for an optimal boundary. We observe that, under some regularity conditions, $p_i \{(1-h_i)R(n-i) - h_i C\}$ is a convex function of p_i . Using Taylor linearization, particularly for group-processing problems with a limited number of steps, the optimal boundary can be solved numerically. The general solution, however, remains to be worked out, possibly using dynamic programming techniques.

5. CONCLUDING REMARKS

In this article, we have discussed several methodological issues related to real-time Dynamic editing. This new data editing method represents an effort to better and more efficiently allocate resources

during the initial data collection and editing process, thus improving the overall operational efficiency and timeliness. Though much work remains to be done, especially regarding the problem of the optimal control boundary, our initial results illustrate the great potential of Dynamic editing in improving the cost-efficiency of the often most costly and time-consuming step in the survey process.

ACKNOWLEDGEMENTS

Claude Poirier, Statistics Canada, Generalized System Methods (Chief)

Operations Research and Development Division,
Statistics Canada

The authors would also like to thank the two referees for their valuable suggestions.

REFERENCES

- Federal Committee on Statistical Methodology (1990). "Data Editing in Federal Statistical Agencies". *Statistical Policy Working Paper* 18.
- Hidiroglou, M.A. and Berthelot, J.-M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". *Survey Methodology*, 12, 73-83.
- Latouche, M. and Berthelot, J.-M. (1992). "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys". *Journal of Official Statistics*, 8, 389-400.