

**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 (vii): Indicators for measuring the quality impact of data editing and imputation

Determining a Set of Edits and Quality of a Database

Supporting Paper

Prepared by Maria M. Garcia and William E. Winkler, U.S. Census Bureau, United States ¹

ABSTRACT

Statistical agencies collect data with the purpose of providing a set of estimates. The collected data must be “cleaned” (data are edited and errors are removed) for reliable and statistically valid publication figures. In this paper we cover methods of creating a set of edits for a set of data and applying them in an efficient, cost effective manner that is intended to maximize quality for a given expenditures of resources. The edits are logical constraints designed to detect errors in the data and are intended to improve the quality of the collected data. If data are edited and erroneous fields are identified and imputed, then they can be used for statistical analyses or business purposes. If errors remain in the data, then it is possible the quality of the data is compromised and data users, including analysts and policy makers will make inappropriate decisions.

I. INTRODUCTION

1. High quality data are needed for modeling and analyses that can affect policy decisions and the understanding of modern society and its economy. In a business, data may be used for tracking customers, evaluating processes, determining where cost reductions may be possible, and data mining for marketing and other business purposes. In government agencies such as regulatory and statistical institutions, data may be collected via censuses and surveys to provide estimates or to track programs.

2. Survey organizations have had a long history of collecting data that is subsequently ‘cleaned’ to remove errors. Much of the early editing of data was performed manually. Many current editing methods are based on more systematic methods in which the manual edit rules are placed in computer code. An edit for discrete data might be where a child in a household cannot be both less than 16 years of age and married. An edit for economic data may examine the average wage paid by a particular company in a particular industry. If the company pays a wage that is much too low or much too high (i.e., in the tails of a probability distribution), we may check whether the average wage associated with the company is in error.

¹ This report is released to inform parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily of the U. S. Census Bureau.

3. Historically, agencies that collect data to be used for statistical, policy, or regulatory purposes have been at the forefront of the field of statistical data editing. They want to assure that published statistics (e.g. aggregates such as totals and rates) are accurate. Statistical agencies developed methods that were often good for cleaning the data. Costs were often high and efficiency low because the agencies did not create systematic methods to determine which of the edit rules were most effective and to minimize the resources needed to achieve a given level of quality.

4. In this paper we cover methods for creating a set of edits. The edits are designed to detect errors in the data and must be applied in an efficient manner with the purpose of improving the quality of the data. Our definition of quality is that the data can be used for its main intended purpose and possibly other purposes not envisioned by the original survey design. We will not cover some issues that affect the quality of data such as how one assures that a list of entities such as customers is complete and unduplicated or how to ask questions in a manner that minimizes respondent errors. We assume that agencies will design a data collection system where it is possible for respondents to answer the questions accurately or for company representatives to enter the collected information accurately in computer files.

5. The outline of this paper is as follows. In Section 2 we provide background information on how editing has been applied. In Section 3 we review existing methods for determining edits. In Section 4 we describe additional methods for determining edits for both discrete and continuous data. In Section 5 we provide ideas on efficiently monitoring the edits along with three metrics for quality. In Section 6 we provide insight into the types of professional skills needed for monitoring and developing specific methods for improving quality. We close with a brief summary.

II. BACKGROUND

6. Most editing methods originated from attempts to assure that keyed data were in a form that could be used for its intended purposes. Statisticians realized that data required editing prior to being used for analyses (Nordhaus (1975), Nordbotten (1963), and more recently Dasu and Johnson (2003)). Many edits could be obtained from the subject-matter experts' knowledge based on their experience reviewing and cleaning the data. Other edits could be obtained by examining statistics in the tail of distributions of data. There are two overall issues of editing. The first is how to determine that an overall set of edits is suitable for cleaning up data sufficiently well so that accurate analyses are possible. The second is how to effectively edit a given database when the set of edits are given. We address only the first issue as significant progress has been made on the second issue.

7. Historically, data editing operations were performed by clerks following a set of if-then-else rules applied sequentially to the data. Clerks may have repeatedly changed fields that had already been corrected if they appear in another failed edit. In this case, editing failed because they lack the implicit information about the original set of edits that may not fail but might fail the imputed record. Fellegi and Holt (1976) demonstrated that these implicit edits are needed for effectively correcting the data. Prior to Fellegi and Holt's work, as individuals corrected records by changing values in fields associated with failing edits, additional edits would fail that did not originally fail. The method has the requirement that the data in each record should be made to satisfy the edits by changing the fewest possible fields. This assures that the corrected record is as close to the original record as possible. The Fellegi-Holt method is global in the sense that if all failing edits (explicit and implicit) for a record are covered (i.e. at least one field in each failed edit is marked to be changed) then by changing the values in the fields associated with the cover we are guaranteed to find a 'corrected' record that fails no edits.

8. The book by De Waal (2003) describes many of the methods of systematic editing that are currently in use in national statistical institutes and survey organizations. The computational issues of large-scale editing have largely been solved (See De Waal (2003), Winkler (1997a), Bruni and Sassano (2001), Franconi et al. (2001), Boskovitz et al. (2003), and Riera-Ledesma and Salazar-Gonzalez (2004)). The issues of software interfaces and training of individuals are still open but are close to being solved in most situations.

III. EXISTING SYSTEMATIC METHODS OF DETERMINING AT SET OF EDITS

9. In a survey setting, editing has often been performed by clerks prior to data being keyed into the computer. The edit rules were often determined by subject-matter experts experienced in similar situations. For instance, items would need to add to a total. A USA state postal abbreviation should correspond to US Postal Service conventions. Data values for certain fields should lie in an interval. Values that are ‘too big’ or ‘too small’ might be flagged as being potentially erroneous. Using experience, the clerk might correct some of the anomalies without contacting a respondent. Although expensive and time-consuming, the clerk might call back a respondent to verify and/or change values and to get values in situations where values were missing.

10. With discrete data, edits were sometimes based on common sense or reasonable expectations. For instance, a child might be required to be at least 15 years younger than the youngest parent. Although there are exceptionally rare situations where this type of edit may not be valid with a particular record, in the overwhelming majority (possibly above 99.99%) of the situations ensuring this edit rule will ‘improve’ the quality of the data.

11. Continuous data, as that collected in economic surveys, are extensively edited. It is known that only ratio and balancing edits are required in more than 99% of economic surveys. In a balance edit, detail items must add to totals. A ratio edit is the requirement that the ratio of two data items is bounded by lower and upper bounds. Ratio edits are straightforward to develop. Typically, we only consider ratios of two continuous variables that are somewhat correlated. Ratios that are far away from the average ratio can be flagged for review. With some survey data (particularly a new survey or a poorly designed one) some of the values of fields in a few of the records will be in error. In those situations, experience has demonstrated that the ratio may be in the extreme tails of the distribution. Such an extreme value is sometimes referred to as an *outlier*. We note that every distribution can have values in extreme tails. With unedited data, the extreme tails of some of the variables will be associated with many more records than would be expected from the number of records associated with the tails of ‘cleaned’ data. Chambers et al. (2000) provide introductory methods for outlier detection and review. The difficulty with these methods is that many (or most) outliers will not be associated with records containing errors. Winkler (1997b) provides methods for converting statistics from the interior of distributions to outliers that can be delineated and reviewed. An open research problem is, “If outliers are ‘corrected’ or changed, then will the resultant database be suitable for accurate analyses and other uses?” A subtle, but related, question is, “If 10% of the values of a variable are systematically increased (erroneously) and still lie in the interior of a distribution, does this type of error cause sufficiently great deviations in aggregates to change the results of statistical tests?”

12. Another type of edit arose as a result of ideas from Exploratory Data Analysis (EDA) of Mosteller and Tukey (1977). Edits that intuitively targeted survey data were introduced by Hidioglou and Berthelot (1986) and refined by Latouche and Berthelot (1992). In the work of Latouche and Berthelot, for each record, an aggregate A_i was defined that was the weighted sum of the continuous variables. Subject-matter specialists determined the weights of individual fields

according to their relative importance. A total $T_A = \sum_i w_i A_i$ was obtained where w_i might be sample weight or some other quantity. Those records associated with T_A that had the highest values of $w_i A_i$ were edited (manually). The Latouche-Berthelot methods were evaluated in comparison with purely manual methods that were often used in economic surveys. Various authors (e.g., Van de Pol and Bethlehem 1997) demonstrated that a thorough follow-up editing of all erroneous units performed by analysts had little effect on aggregates of the form T_A . This is because errors in the aggregates are often due to errors in fields associated with a few important units. Granquist and Kovar (1997) provide a survey of selective editing methods such as those of Latouche and Berthelot (1992). The intuitive idea of selective editing with aggregates of the form T_A is that it helps control the amount of editing. Reporting units are ranked and prioritized for review. Units that contain influential errors with a larger effect on the final estimates are high on the ranked list and are manually edited. A thorough review of low ranked units will have a negligible effect on the final estimates and no further improvements are obtained by manually reviewing them. These units are either not edited or edited using an automated system.

13. Garcia and Thompson (2000) pointed out that a Fellegi-Holt system could automatically edit a set of survey data in 24 hours that 12 analysts took six months to edit. In this case over-editing may be a concern as the analysts changed three times as many values in fields as the Fellegi-Holt system. There is anecdotal evidence (Kovar (2005) private communication, Lyberg (2005) private communication) that unsystematic changes by analysts may actually reduce the quality of a database in some situations. These situations point out two issues that need to be addressed. The first is that we need to control the amount of editing performed by analysts. Without systematic control, the clerks may over-edit, possibly to the extent of reducing the quality (accuracy here) of various aggregates obtained from a database. The second is that we would like to identify edits that are the most efficient in the sense that they minimize the number of records that are reviewed and maximize the number of changes (i.e., errors) that are found in the edited data. We note that the changes are only those that bring the values in a record closer to some underlying ‘truth.’

14. In this section we presented existing methods for determining a set of edits. The edits are often based on analysts’ experience, common sense, and reasonable expectations or they can be derived to ensure that certain aggregates are as accurate as possible. Winkler (1999, 2003) and De Waal (2003) suggest that systematic methods such as Fellegi-Holt be used to automatically edit all records. Analysts would only review and clerically change those records that have a potentially significant effect on publication totals or on some important aggregates. To assure that all data satisfy the edit rules, heuristic algorithms would be used for the final ‘correction’ of data and have little effect on most aggregates.

IV. NEW IDEAS ON DETERMINING A SET OF EDITS

15. The primary purpose of this section is to describe methods for systematically editing collected data to ensure that a database can be used for statistical analyses. We break the methods according to discrete and continuous data. We also discuss imputation.

A. Discrete Data Edits

16. In this sub-section, we will consider determining a set of edits for discrete data in situations where a survey form is well-designed and questions are asked effectively. If the survey form is not well-designed, then there can be large systematic errors in the data that are often quite

straightforward to identify. We consider the situation where errors are due to transcription, keypunch, and isolated misunderstanding of questions.

17. To describe the situation more effectively, we slightly digress to describe how discrete data $X = (X_1, X_2, \dots, X_n)$ might be used in a loglinear model. In loglinear modeling (i.e., Agresti 2007; Bishop, Fienberg, and Holland 1975), we may determine that all 3-way interactions and no higher order interactions yield a model that provides a good fit to the data. The 3-way interactions are the marginal counts or marginal probabilities

$$P(X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, X_{k_3} = x_{k_3}) \quad (1)$$

where X_{k_1}, X_{k_2} , and X_{k_3} are any three fields in a set of n fields X_1, \dots, X_n that can be easily computed from the original data X . It is known that the largest probabilities of form (1) determine most of the accuracy of the good fit provided that the very small probabilities of form (1) do not in total account for a moderate amount of the overall probability.

18. In analyzing the data, an experienced statistician might realize that one or more of the largest probabilities of form (1) are too low in comparison to the equivalent probabilities from a comparable alternate source. If this is true, then a number of the very small probabilities of form (1) may be nonzero due to keypunch error that has changed one or more values in fields. To find these small ‘erroneous’ probabilities, we might first enumerate and examine the patterns $\{X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, X_{k_3} = x_{k_3}\}$ associated with probabilities in the range [0.001, 0.01]. If something looks unusual (and we possibly verify with a subject-matter expert), then create the appropriate edit. We might then consider probabilities in range [0.0001, 0.001]; if some are in error, determine whether it is suitable to use them. The intuition is that we want as few ‘edits’ as possible and that many of these small ‘error-probability’ situations may not seriously affect overall use of the data.

19. A second way of determining a potential set of edits is to have two data sources that represent similar subpopulations of a larger population, in which the two subpopulations have several variables in common, and for which one source has been edited. In this situation, we can tabulate probabilities of the form $P(X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, X_{k_3} = x_{k_3})$ for any three (or two or four variables) that are common to the two files. Because the subpopulations are representative of a larger population, the probabilities from the set of tabulations should be similar in most situations. We consider the situation where the specific probability for a pattern of the form $\{X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, X_{k_3} = x_{k_3}\}$ from the unedited data source is much higher than the specific corresponding probability from the edited data source. If the unedited-file probability is nonzero and the edited-file probability is zero, then data pattern $\{X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, X_{k_3} = x_{k_3}\}$ may be a possible edit and needs to be verified with subject-matter specialists. If suitable a priori information is available (such as the specific edits that were used in the edited data source), then it would be easier to take the appropriate specific edits that were used for the edited data source.

B. Continuous Data Edits

20. In this case we assume that the survey collects data that are multivariate continuous, $X = (X_1, X_2, \dots, X_n)$. We consider pairs of variables X_i and X_j that have pairwise correlations ρ_{ij} above a certain point (e.g. $\rho_{ij} \geq 0.4$). When the correlation of fields X_i and X_j is low, the corresponding ratio of the two variables will have a wide spread and may not provide useful

information. If the two variables are not somewhat correlated (i.e. $\rho_{ij} < 0.4$), then one has almost no value in predicting the other. Looking at these relationships between the variables can provide information to use for editing or later during imputation. For convenience, we assume that the variables X_i and X_j are positive (except when missing). We look at upper and lower tails (say 2% and 98%) of distributions of the ratios of the variables as they may delineate possible erroneous records. We can then do follow-up to determine whether some of these records associated with the tails of distributions are in error. The tails of distributions of ratios may point to a pair of potentially erroneous field but do not identify which of the two fields is incorrect. We can also use subject-matter knowledge of the expected relationships between the fields to identify errors. By considering the relationships of the potentially erroneous fields with other data fields, we could possibly determine which field should be corrected. A good corroborating factor is if X_i is in the tail of distribution of ratios with X_j and X_j is in the tail with some other variable X_k . That is, if X_j is in the tail of distributions with two or more variables, then it is more likely to be in error.

21. A conjecture is that this type of procedure will identify most of the variables of a multivariate record that are in error. We do not need to look at full multivariate distributions. If a variable X_j is not correlated with any other variables in the set $\{X_1, \dots, X_n\}$, then there is no way for the variable X_j to be in any effective ratio edits or any other edits that are based on the data fields in the existing database.

C. Imputation

22. In the previous subsections we presented strategies for determining more edits for both discrete and continuous data. The purpose of applying the edits is to identify erroneous data items that must be imputed. At the crudest level, we can always substitute in one of the values of a variable that will cause a record to no longer fail edits. We can do these using sequential Fellegi-Holt methods of Chen et al. (2003) or Garcia (2004). Few imputation systems assure that records satisfy edits (Winkler 2003).

23. At present, it is not clear how we will substitute in values that also preserve probability distributions. For discrete data, Winkler's (2003) procedure is likely to yield suitable probability distributions (from which to draw values) for small situations (eight or less variables). Although the methods are theoretically valid for arbitrarily large situations, the methods are not presently computationally tractable. It is not clear how to effectively extend the probability-distribution computational procedures with heuristic or other methods. Winkler (2003) provides two computational simplifications that may be suitable in some situations. More recently, Winkler (2008) developed models and exceedingly fast algorithms that assure that imputation maintains joint inclusion probabilities and create records that always satisfy edit constraints. For continuous data, we may need to break up each continuous variable into a set of discrete subdomains. In this continuous-to-discrete situation, we could compute the overall probability distributions. It seems that this approximate procedure will be suitable for practical applications. Bruni (2004) provides specific methods for the continuous-to-discrete conversion.

V. MONITORING QUALITY

24. In this section, we describe a strategy for efficiently monitoring the edits and minimizing the resource levels needed for achieving a given level of quality. Although understanding of quality is often subjective for users of a data file, we wish to provide some quantification. For instance, an individual may state the data needed for running a business has suitable quality for

day-to-day operations but needs to be improved for secondary uses. Often, some elementary checks will show that the data also need improvement for the primary purposes. The strategy that we describe for is intended to enhance and combine a number of existing methods. In application, all of the existing methods for improving the quality of data must have components that are quite specific for a particular type of data and use. These very specific methods are unlikely to be suitable for other types of data and uses. Our aim is to provide a framework for monitoring and/or improving the quality of data within a given level of resources while providing the most efficient methods for the 'clean-up' of the data.

25. We provide three (partial) metrics for quality. The first metric is the *number of edits* that are used in 'correcting' the data. We would like to believe that, if ten edits are initially used, then fifteen edits are better. The second metric is the *precision of an edit*, where precision is defined as the proportion of records delineated by an edit that are actually in error. Granquist and Kovar (1997) refer to query edits as those that are not always associated with errors in the data. If we use the tails of distributions of continuous variables to identify items that have a high probability of being erroneous, then we would like a high proportion of the edit-delineated records to be in error. The third metric is the *proportion of records that are affected by an edit*. For instance, if 0.0001 percent of records fail an edit, then we may not need to apply this edit rule to the data. There are subtle issues associated with proportion of records. With discrete data where populations are not skewed, it may be easy to include an edit for situations where a wife is 50 or more years older than her husband. If the inclusion of such edits adds considerable complication to the overall edit system, then we may not include them. With continuous data, we may wish to look in the tails of a distribution if it is associated with the largest companies that affect many statistics the most. In this case we may also wish to include the less precise edit. For instance, if an edit delineates many records, only about 10% of the delineated records are in error, and the changes yield significant improvements to key aggregates in a database, then we may wish to include the less precise edit. The advantage of knowing those edits that are delineating records and achieving improvements in aggregates is that we can much more easily look for alternative edits that delineate most or all of the set of erroneous records and are more precise.

26. In a survey situation, most of the resources will be needed for determining additional edit constraints. With discrete data, it may be possible to try additional analyses (loglinear and otherwise) or look for additional association rules that may be in error. Any additional error conditions can be added to the existing set of edits. Updating the existing edit tables and running the edit/imputation programs uses very minimal resources.

27. With continuous data, the distribution of ratios of pairs of fields may change from time period to time period. In each situation, the bounds (upper and lower) in the edits will need to be updated in the edit tables. Additional outlier detection methods might be used to determine any multivariate outliers. If multivariate outliers are determined, then it might be possible to determine whether any new ratio edits will delineate the same outliers that the multivariate-outlier-detection methods have determined.

28. Analysts who use the data in certain analyses may identify additional outliers and edits in an ad hoc fashion. If the additional outliers are considered serious, then additional resources will be needed for updating the entire edit/imputation system. The amount of computation in a Fellegi-Holt system grows at an exponential rate with the number of edits. If we wish to add m new explicit edits to an existing set of n edits, then worst-case computation will increase from $O(2^n)$ to $O(2^{n+m})$ with some of the records. If an edit does not affect many records and leaving the errors in records will not seriously affect important aggregate estimates, then it may be best to leave the errors uncorrected.

29. An open research problem is how to quantify the quality of a statistical database (even for one set of analytic uses). If a database (particularly one with continuous economic data) has been through substantial editing, then is it possible to quantify the improvement due to the edit/imputation? In some situations a database that is routinely subject to extensive edit/imputation is later discovered to contain additional (possibly severe) errors. The errors may be discovered via a routine analysis of the data or via comparing certain aggregates computed with the data with external data aggregates. An issue is “If severe errors are discovered in a database, then why was the database considered (subjectively) of reasonable quality and what effect do the newly discovered errors have on previous analyses and decisions based on the data?”

We suggest the following minimalist strategy for improving quality in a database.

1. Provide very succinct documentation of the purpose of a system, where files are located, and the metadata of a database.
2. Determine a list of the key aggregates (statistics) that a system must produce.
3. Describe and precisely quantify the procedures used to improve the accuracy of the aggregates.

Key aggregates may consist of a few totals of variables. They may also consist of the means and covariances of a set of variables if the set is used in regressions. For more sophisticated analyses, the aggregates may take the form of sums and moments (Moore and Lee (1998), DuMouchel et al. (1999)) used in computing the likelihoods associated with models of the data.

A. An illustrating example using artificial data

30. In this subsection we present an example using continuous data. Our goal is to illustrate that the tails of the distributions of ratios can be used to determine a set of edits that is suitable for cleaning up the data sufficiently well for producing accurate estimates. We use available artificial ‘raw’ numeric data and corresponding ‘cleaned’ data from the Annual Survey of Manufactures (ASM). The ASM measures manufacturing activity that includes employment, payroll, fringe benefits, product shipments, capital expenditures, and total inventories. This survey is the only source of comprehensive data on the manufacturing level of the American economy. Each record contains data for 17 numerical fields. Every pair of fields is subject to a ratio edit. In addition there are two additive edits for verifying totals for salary and wages and total employment balance to the sum of their respective details.

31. To produce the artificial raw data we started from a subset of records with no edit failures and randomly introduced errors into every record in the file. This test deck contains an unrealistically high proportion of records with a large number of errors by design. The error generation mechanism is not designed to simulate actual non-sampling errors but to induce errors in such a way that a randomly chosen explicit or implicit edit fails. For every record we repeated the generation of artificial errors multiple times to obtain an artificial data file consisting of 9116 records.

32. To simulate what happens in a real production situation we kept 20% of the erroneous records in our test data. In a real production setting only a proportion of the records are in error. If a large proportion of records (say $\gg 20\%$) are edit failing records then the pairwise correlations of the variables’ will be small (say $\rho < 0.4$) and a given variable will have little predictive value over the others. We used the SPEER editing system (Draper and Winkler (1997), and Garcia (2004)) to produce two files of ‘cleaned’ data; one file using the corresponding set of edits and another file using the tails of the distributions of ratios to identify errors in the data. In Section 4.2 we conjectured that it is possible to identify most errors in the data by looking at the tails of distributions of ratios of correlated fields. The criteria for identifying edit failures (ASM subject matter edits vs. edits based on tails of distributions of ratios) are different. Thus we do not expect

to completely delineate all errors in the data. However, we would expect the cleaned data edited using the tails of distributions of ratios could be effectively used for the same purposes and analyses as the data edited using the subject matter prescribed edits. The cleaned data should be usable for calculating certain aggregates like totals and simple statistics like means and medians, and regression parameters if the data set is used for some regression analyses. Figure 1 examines the distributions of estimated totals for all fields using both sets of cleaned data. The plot illustrates that we are able to use the data cleaned using the tails of distributions of ratios to approximate the distribution of aggregated totals. The differences in the estimated totals using both sets of cleaned data are small with the only exception of field X_{16} . It is beyond the scope of this paper to explain why there is discrepancy in the estimated totals for this field. Our goal is not to evaluate the procedure as applied to the ASM data but to illustrate that it is possible to detect most errors using the upper and lower tails of the distribution of ratios as a criteria for identifying errors in the data. The data can then be cleaned in such a way that it can be used for statistical analyses and for estimating certain aggregates like totals. [Note: For additional examples see our larger research report, to appear.]

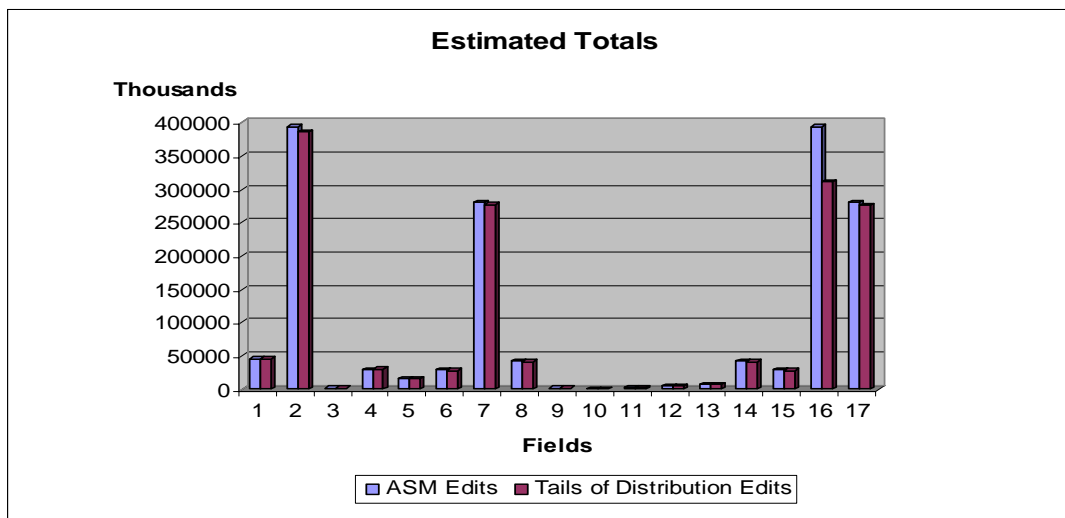


Fig 1: Estimated totals using ASM subject-matter edits and tails of distribution edits

VI. NEEDED TECHNICAL SKILLS

33. In this section we talk about the professional skills needed for designing and running an edit/imputation system, systematically monitoring the edits, and developing specific methods for improving quality in a database. A statistician or other analyst with suitable skills for running straightforward software could do most (or all) of the edit/imputation with little (or no) assistance from subject-matter experts. In the easiest database situations, it seems plausible that the data from the simplified, one-person procedure are of higher quality than the data produced via ordinary methods (i.e., classical manual editing or application of many subject-specific if-then-else rules.) In other situations, subject-matter expertise may facilitate the statistician's improvement of the data. In no situation does it seem likely that nonsystematic review and 'correction' by analysts would yield a final database of comparable quality to the database created by the statistician that employs the simplified, one-person procedure to improve the data.

34. Substantial skill is needed for creating a basic Fellegi-Holt type of system for editing and imputing either discrete or continuous data (Winkler 1999, Winkler and Hidiroglou 1998). If the basic editing software has been created, then the skills needed for running the software and

performing analyses are often less than those for analyzing messy data using SAS. The development and utilization of quality-improvement methods will possibly require an institutional change in how an agency maintains a database or a set of databases.

VII. SUMMARY

35. In this paper we provided background on existing methods for determining a set of edits of a statistical database. These previous methods are often based on subject-matter expertise. We also presented ideas on easily implemented methods for determining additional edits for a given set of data along with strategies for efficiently monitoring the edits and applying them in an efficient manner with the goal of maximizing the quality of the database. The idea is to significantly reduce the number of edits and resources needed for ‘correcting’ (i.e. edit/imputing) data.

REFERENCES

- Agresti, A. (2007), *An Introduction to Categorical Data Analysis (2nd Edition)*, New York, NY: J. Wiley.
- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.
- Bruni, R. (2004), “Discrete Models for Data Imputation,” *Discrete Applied Mathematics*, 144, 59-69.
- Chambers, R., Hentges, A., and Zhao, X. (2004), “Robust Automatic Methods for Outlier and Error Detection,” *Journal of the Royal Statistical Society, A*, 167 (2), 323-339.
- Chen, B.-C., Thibaudeau, Y., and Winkler, W. (2003), “A Comparison Study of ACS If-the-else, NIM and DISCRETE Edit Systems using ACS Data,” UNECE Statistical Data Editing Work Session, Madrid, Spain, <http://www.unece.org/stats/documents/2003/10/sde/wp.7.e.pdf>.
- De Waal, T. (2003), *Processing of Erroneous and Unsafe Data*, ERIM Research in Management: Rotterdam.
- Dasu, T. and Johnson, T. (2003), *Exploratory Data Mining and Data Cleaning*, Wiley-Interscience: New York.
- Draper, L., and Winkler, W., (1997), “Balancing and Ratio Editing with the New SPEER System,” *American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods*, 582-587.
- DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999), “Squashing Flat Files Flatter,” *Proceedings of the ACM Knowledge Discovery and Data Mining Conference*, 6-15.
- Fellegi, I. P., and Holt, D. (1976), “A Systematic Approach to Automatic Edit and Imputation,” *Journal of the American Statistical Association*, 71, 17-35.
- Garcia, M. (2004), “Implicit Linear Inequality Edits Generation and Error Localization in the SPEER Edit System,” *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM.
- Garcia, M., and Thompson, K. J. (2000), “Applying the Generalized Edit/Imputation System AGGIES to the Annual Capital Expenditures Survey,” *Proceedings of the International Conference on Establishment Surveys, II*, 777-789.
- Granquist, L., and Kovar, J. (1997), “Editing of Survey Data: How Much is Enough?” in Lyberg, L., Biemer, P. Collins, M., De Leeuw, E. Dippo, C., Schwartz, N. and Trewin, D. (eds), *Survey Measurement and Process Quality*, John Wiley and Sons, 415-435.
- Hidioglou, M.A., and Berthelot, J.-M. (1986), “Statistical Editing and Imputation of 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 Business Surveys,” *Journal of Official Statistics*, 8 (3), 389-400.

- Moore, A. W., and Lee, M. S. (1998), "Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets," *Journal of Artificial Intelligence Research*, 8, 67-91.
- Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.
- Nordbotten, S. (1963), *Automatic Editing of Individual Statistical Observations*, Statistical Standards and Studies, Handbook No. 2, United Nations: New York.
- Riera-Ledesma, J., and Salazar-Gonzalez, J.-J. (2004), "A Branch-and-Cut Algorithm for the Error Localization Problem in Data Cleaning," technical report, Universidad de la Laguna, Tenerife, Spain.
- Van De Pol, F., and Bethlehem, J. (1997), "Data Editing Perspectives," *Statistical Journal of the United Nations ECE*, 14, 153-171.
- Winkler, W.E. (1997a), "Set-Covering and Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 564-569 (also available as Statistical Research Division Report RR98/01 at <http://www.census.gov/srd/papers/pdf/rr9801.pdf> .)
- Winkler, W. E. (1997b), "Problems with Inliers," paper presented at the European Conference of Statisticians, October 14-17, 1997, Prague, Czech Republic, http://www.unece.org/stats/documents/1997/10/data_editing/22.e.pdf .
- Winkler, W. E., and Hidirolou, M. (1998), "Developing Analytic Programming Capability to Empower the Survey Organization," Statistical Research Division Report RR98/04, <http://www.census.gov/srd/papers/pdf/rr9804.pdf> .
- Winkler, W. E. (1999), "The State of Statistical Data Editing," in *Statistical Data Editing*, Rome: ISTAT, 169-187 (also available at <http://www.census.gov/srd/papers/pdf/rr99-01.pdf> .)
- Winkler, W. E. (2003), "A Contingency Table Model for Imputing Data Satisfying Analytic Constraints," *American Statistical Association, Proc. Of the Section on Survey Research Methods*, CD-ROM, also research Report SRS 2003/07 at <http://www.census.gov/srd/papers/pdf/rrs2003-07.pdf> .)
- Winkler, W. E. (2008), "General Methods and Algorithms for Modeling and Imputing Discrete Data Under a Variety of Constraints," Statistical Research Division Report RRS2008-08, <http://www.census.gov/srd/papers/pdf/rrs2008-08.pdf> .