WP. 52
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (v): Quality indicators and quality reporting

MODELING AND ANALYSIS WITH DATA

Supporting Paper

Submitted by the U.S. Census Bureau, United States[1]

ABSTRACT

Survey data are sometimes collected for a single purpose such as providing a set of estimates. If the microdata are sufficiently clean (without error), then it is possible for the data to be used for a variety of analytic purposes. Most analytic purposes involve building a model based on the data. This paper describes modeling in terms of data quality issues. Anomalies in the aggregate quantities represented in the model may delineate subsets of the microdata that need to be edited and imputed.

I. INTRODUCTION

1. Data are often used for a variety of purposes. For instance, continuous microdata representing energy inputs and outputs may be used in an economic model. An econometrician may create a regression-type model describing the overall quantities and costs associated to the inputs of energy-related products such as natural gas, petroleum products, electricity, or coal and the outputs of manufactured goods and chemical products. The model may represent a group of companies in a particular industrial sector. The same data may be used for simpler analyses in which the quantities of various inputs are correlated with the quantities of various outputs. Or, most simply, the data may be used to create certain aggregates such as totals.

2. Discrete data may be collected in a census or a population survey. At a simple level, the data may be used to create tables. Some tables might describe the number of people in various age and sex categories within geographic regions such as U.S. States or municipalities. Simple analyses that are microdata-dependent may yield association rules (Agrawal and Srikant, 2000, Hastie et al. 2001) that are based on conditional and joint probabilities of occurrence of certain combinations of values of variables. More sophisticated analyses may use loglinear models.

3. All analyses are dependent on certain aggregates that are computed from the microdata. Different analyses are dependent on different sets of aggregates. Each set of aggregates places restraints on the microdata. If certain aggregates can be determined to be in error, then some of the records corresponding to the aggregates may be in error. If a record is in error, then the values of certain

[1] Prepared by William Winkler william.e.winkler.@census.gov and Maria M. Garcia maria.m.Garcia@census.gov
This report is released to inform parties of research and to encourage discussion. The views expressed on methodological and technical issues are those of the author and not necessarily those of the U. S. Census Bureau.

variables in the record may be erroneous. For instance, the data associated with a company record may provide total number of employees and total payroll for which the ratio (average wage) is far higher than the wages known to be paid in a particular industry. The record might be 'corrected' by determining which variable (or variables) need to be changed. The changed value(s) presumably yield a record that is acceptable.

4. The process of 'correcting' a record is known as edit/imputation. In the *edit* process, values of variables (fields) in records are located that may be in error. The erroneous value in a variable may be replaced by a value that is known to be acceptable. Replacing the value of a variable or putting a value in a blank variable is known as *imputation*. Subject matter experts reviewed and changed records via clerical review in early edit systems. The subject matter experts often had to learn some of their expertise from individuals who analyzed the overall data. Determination of errors was via call-backs and follow-up with the individuals who had filled-in the questionnaires at individual companies. Enumerators often re-interviewed individuals to 'correct' some household surveys.

5. As computers became widely available, statistical agencies and others put the edit/imputation rules in computer programs. The first general issue was whether the subject matter analysts had edit/imputation rules that would find most of the errors in the data and, possibly, whether the imputations were appropriate. The second issue was that, as variables associated with failing edits were changed, additional edits that had not previously failed would fail. Fellegi and Holt (1976) solved the problem associated with the second issue. They demonstrated that implicit edits (those edits logically derived from explicitly defined edits) were also needed for 'correcting records.' Further, Fellegi and Holt demonstrated how the edits could be placed in easily maintained tables rather than hundreds or thousands of if-then-else rules. They did not deal with the computational issues of fast integer programming methods that still remain a research issue for surveys having hundreds of edits or millions of records.

6. Determining an appropriate set of edits for a given data source is still an open problem. The basic set of edits suggested by the analysts may be incomplete. Exploratory data analysis (EDA) may suggest different edits (e.g., Hidiroglou and Berthelot, 1986; Des Jardins, 1998; Des Jardins and Winkler, 2000). Some new edits may be found via selective editing methods (e.g., Latouche and Berthelot, 1992) that look at records that affect certain aggregate quantities the most. Further, various outlier detection methods (Kosinski, 1999; Rocke and Woodruff, 1996) may be used. In all situations, edits represent unusual situations in a given data record or in a group of records. In many situations, edits are determined by the relationship of a particular record to an aggregate. For instance, in the payroll-employment example, a very high or low ratio in relationship to the average ratio in an industry may represent an error in the data record. The average is the aggregate.

7. When analysts create models with microdata, there are several issues that need to be considered. The first is whether the survey frame (list of entities) associated with the microdata is complete, unduplicated, and, where applicable, has appropriate quantitative data such as a prior year's total sales associated with it. The second is whether the microdata are complete (have no blanks) and have appropriate values associated with individual records. In this paper, we only deal with the second issue in a particular context. Our context is that an individual performs an analysis on the data. In some situations, the individual may do an analysis that corresponds to purposes for which the data are collected. Because data is a valuable resource, the individual may do analyses that do not correspond to the reason the data were collected. In all situations – even with data that are edited and imputed – the aggregates in the model or auxiliary analyses performed to validate model assumptions may yield potential errors in the microdata.

8. It is our belief that collected data, as an ideal, should be suitable for multiple analyses. To assure this, the cost of 'correcting' the data should be minimized. Increasingly, individuals (Granquist and Kovar, 1997) have shown that edit/imputation methods are not applied in an efficient manner. For instance, van de Pol and Bethlehem (1997) have shown that only 10-20% of the changes applied to survey microdata may be needed to assure reasonable accuracy of the main totals. More dramatically, Garcia and Thompson (2000) demonstrated that twelve analysts needed six months to edit and impute a

large, economic survey whereas a Fellegi-Holt system required 24 hours and changed 1/3 as many fields. Although we do not advocate editing data solely using automatic methods, we suggest that an automatic method might be used to initially process all microdata records. Subject matter experts might additionally review records that affect aggregate totals the most. Software as in Greenberg and Petkunas (1990) would assist the clerks to enter new values for variables in records that satisfy edits. Where edits cannot be entirely satisfied, the clerks might enter override flags to indicate that certain variables in an individual record can no longer be edited. The assumption is that the clerk, through a call-back or other means, has found values for certain fields that correspond to an individual entity's 'real' values even though the values do not satisfy edits.

9. The outline of this paper is as follows. The second section describes methods for edit/imputation more fully. It provides some elementary examples from existing literature. In the third section, we describe how someone might analyze a set of data. We relate the aggregates from some of the analyses to current edit/imputation methods and suggest where certain types of analyses may suggest additional types of edits and imputations. The fourth section provides discussion. The final section is concluding remarks.

II. BACKGROUND AND ELEMENTARY EXAMPLES

10. In this section, we first provide some examples of elementary edits. We next describe how certain analyses of data are performed. The idea is that the analyses are better done when the data are 'corrected' via edit/imputation procedures. If anomalies arise during an analysis, then it may be possible to delineate additional edits that, in turn, may yield additional records that may need edit/imputation. Throughout this paper, we will also assume that a true underlying value of each variable in a record exists. This means that a survey questionnaire may ask questions that correspond to some underlying reality and can be reasonably answered (with a given set of time/resources). Errors may be random such as from transcription or keypunch. They may also be systematic as when a subset of the population in a file has reported in the wrong units (say, cents instead of dollars). In some situations where the value in a variable in a record is in error, a correct replacement may be obtained via a call-back to a company or via a re-interview.

11. Given a specific set of data, a fundamental question is "How do we determine a suitable set of edits?" The purpose of edits and imputations is to produce 'corrected' data records that are closer to the 'true' values. If the data are edited and imputed using reasonable methods, we would hope that the data are suitable for one or more analyses. By being used for an analysis, we mean that the analysis, however simple, can represent some type of aggregates that are useful. For instance, in an employment database, we might wish to get the counts of individuals in a company or set of companies in various ranges of income. We would hope that the database has sufficiently accurate records so that the numbers of individuals in the income ranges are close to the underlying real values. A second question is "If we are performing an analysis, how do we determine whether the records in the database are sufficiently accurate for the analysis?"

Elementary Examples

12. The first example deals with discrete data such as data collected in demographic surveys.

Example 1. Edits of Demographic Data

13. A census of population collects data that represents a set of households and the persons that reside in the households. Within the household, one person is designated as head-of-household. For each person, the survey may collect variables relationship to head of household, sex, marital status, age, date-of-birth, and race. If the relationship is son, then the sex should be male. If the age is less than 16, then the marital status should be single. If the ages of the parents (head of household and spouse) are each approximately twenty-five, then the age of the child should be considerably less (say, twenty years less)

than the ages of the parents. If the relationship is son and the sex is female, then either the value of son should change or the value of sex should change. Similar changes would need to be made with the other edit conditions. If the age of the wife were twenty or more years greater than the husband, then subject matter experts might decide that an edit should require that the age of the wife cannot be twenty or more years greater than the husband. In many situations, if we have analyzed existing data sources, we might not include an edit for age of wife twenty-or-more greater than husband if the error rarely occurs (say one time in hundred thousand).

14. In a simple tabulation of information from the census age ranges of children within households against age ranges of parents, additional edits may be determined. For instance in situations where the child is forty or more years younger than the mother, subject matter experts may decide that an edit is needed. Although it is possible for a mother to be forty or more years older than a child, the analysts, based on subject matter expertise, may decide the overwhelming majority of situations (above 95%) where children are forty or more years younger than their mothers are associated with errors that need editing.

15. With these simple error conditions, straightforward tabulations will show errors. These types of errors, if left in files, will yield substantial errors in any loglinear analysis that individuals perform on files. The errors might make it impossible to perform association rule mining (Hastie et al., 2001) in which joint and conditional probabilities for a few variables (say, four or less) are computed.

Example 2. Edits of Continuous Data

16. Most edits of continuous economic data consist of ratio edits and balance equations. In balance equations, items must add to totals. Both these types of edits are special cases of linear inequality editing. De Waal (2003) describes a system for linear inequality editing based on Fourier-Motzkin equation solving. Draper and Winkler (1997) and Garcia (2004) provide edit methods for the much simpler situation consisting only of ratio edits and balance equations that are far faster than the general linear inequality methods. If, say, we are editing a large number of companies ($i = 1, 2, \ldots, N$) that report total payroll P_i and total number of employees T_i, then we may wish that the ratio P_i/T_i to be in an interval [L, U].

17. The intuition is that the ratio P_i/T_i cannot be too high or too low for a particular industry in a particular year. If we consider the overall distribution of P_i/T_i from a prior time period, then the bounds L and U may be obtained by adjusting upwards (or downwards) for the current time period. The intuition is that if the ratio P_i/T_i is above U and below L (i.e., in the tails of the distribution), then there is a potential edit failure. If a ratio Pi/Ti fails the edit, then we may need to change P_i or T_i so that the ratio lies between L and U. What can happen in practice (particularly in the first year of a survey) is that there are greater numbers of reporting errors or keying/transcription errors that change the observed reported value in the computer file of P_i or T_i significantly from the true values associated with a company. The very largest deviations may cause the ratios P_i/T_i to be much too large or much too small.

18. The bounds L and U can be separately found for particular industries. Within an industry (according to North American or European Union coding), the bounds may target subsets of companies such as the largest and the smaller ones because larger companies may have different characteristics (in terms of edits) than the smaller ones. The bounds L and U can also be determined quite quickly using survey data from the current time period after a sufficiently large amount of data has been collected.

Example 3. Outlier Detection Methods

19. Although ratio editing is a simple form of outlier detection, more sophisticated (and more difficult to implement) methods for outlier detection may delineate subsets of companies for which a higher proportion of records actually contain errors. Assume our survey data X has individual records of the form $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ for $i = 1, \ldots, n$. For convenience, assume that we are able to transform the data X into a form that is approximately multivariate normal distribution. We also denote the

transformed data by X. Then, Mahalanobis distance d(Xbar, X_i) may be used as an outlier detection method. Determining the upper bound U for the Mahalanobis distance may be straightforward. Only general linear inequality edits (De Waal, 2003; Riera-Ledesma and Salazar-Gonzalez, 2004) can be used for determining the minimal number of fields to change so that the resultant changed record satisfies edits.

Example 4. Inlier Detection Methods

20. Various authors have observed that certain values of variables or sets of variables that are in error will still not be in the tails of distributions. For instance, with certain surveys, individuals observed that certain companies reported the exact same values for variables in prior time periods. An easy edit is to check a company's current reported values against prior year's report to determine whether any item's values are identical for the two time periods. Winkler (1997) suggested using different sets of variables. For instance, a ratio like total payroll to total number of employees P_i/T_i might not be an outlier. If P_i or T_i, were edited against other variables x_{i1}, x_{i2}, and x_{i3}, it is possible the resultant statistic (possibly a simple ratio) would yield a new outlier. The new outlier would cause either P_i or T_i to be changed.

21. There are several observations that we can make based on these four examples. All of the methods (with the possible exception of the outlier detection methods of Example 3) use simple ideas to detect possible error situations using elementary combinations of variables. With continuous variables, we can think of a ratio of two variables as a simple statistic for which we analyze its distribution. We may consider some or most of the correlated pairs of variables and consider intervals outside of which the ratio is a type of outlier that may yield a value or values that are in error.

22. How do we determine sets of variables that may yield information suitable for edits? Winkler (1997, Example 2.d.) considered a situation in which variables X_1, X_2, and X_3 were not highly correlated with a variable Y. In such a situation the ratios X_1/Y, X_2/Y, and X_3/Y would not yield moderately high information for editing. He observed that a linear combination of variables $Z = a X_1 + b X_2 + c X_3$ had much higher predictive ability of Y and could be used to obtain an edit. With discrete data, the only methods of determining sets of variables to be included in edits are using knowledge from subject matter experts or from individuals who have extensively analyzed similar data.

23. The main observation is that simple combinations of variables are likely to yield the most appropriate set of edits. Many combinations of larger sets of variables are unlikely to yield aggregates that are suitable for edits. By *aggregates*, we mean combinations of variables that can be compared effectively and used to create an edit. Any analysis will yield many aggregates. For instance, a multivariate regression analysis will yield regression coefficients, predicted values, residuals, variances of coefficients, and other aggregates. In loglinear modeling (Bishop et al. 1975), iterative proportional fitting is used to fit a set of lower order interactions to the best fitting approximate answer. The approximate answer is typically a more parsimonious model than the main model that involves interactions of all variables. Each interaction restraint is obtained by fitting to marginal subsets of variables. The iterative fitting proceeds through all restraints multiple times until sufficient accuracy of the fitted solution is obtained.

24. The more parsimonious model sometimes yields better understanding and interpretability of the model and data. Analogously, in a regression model, using the best six variables instead of all ten variables may yield better understanding. The advantage of having ten variables instead of six is that the redundant four variables may be used to edit/impute the best six and improve overall quality of the data (at least for a particular regression analysis).

III. OVERALL METHODS

25. Within the framework of methods to be used on arbitrary types of data, we need to determine the edits that might be needed for a given analysis or set of analyses. If we think about how we create models, then we need to think about what errors in what variables might affect the analysis. With

continuous data, a good overall strategy is to look at the distributions of pairs and groups of variables. With pairs of variables, we may only use pairs that are correlated. With groups of variables, we may determine a subset of variables that predict one of the variables. The outliers in the distributions are good starting points for delineating records that may need editing. Experience in surveys has shown that a poorly designed survey or a first-time survey will yield a much greater amount of error than a well-designed survey that has been implemented more than once.

26. With an employee database containing name, address, date-of-birth, age, sex, salary, years employed by company, department, and other information, we may wish to break out salaries by ranges and by sex or age. A simple check of printing sex and first name may show where the sex code (i.e., 'M' or 'F') is in error or missing. If the sex of 'Susan Smith' is given by 'M' then there is a likely error. If the sex code is given by (0-missing, 1-male, 2-female), then values above 2 are in error. If the number of years employed is slightly less or greater than the age, then age or years employed are likely to be in error.

27. Editing a single field can be straightforward. For instance, the two-character U.S. Postal State Code can only take certain values. If the salary range of a given employee in a specific job category is known to be in the interval $[L_B, U_B]$, then any salary outside the interval is known to be in error.

28. With continuous data, editing multiple fields simultaneously may be straightforward. In the simplest situation, subject matter specialists may supply a set of ratio edits (possibly with upper and lower bounds) and balance equations that must be satisfied. More generally, we could do any exploratory analysis of the data $\mathbf{X} = (X_{ij})$ where $\mathbf{X}_i$ represents the continuous data in a record associated with an entity. If we let $X_{i(j)}$ represent all the variables except the j^{th} column, then we may use $X_{i(j)}$ to predict the j^{th} values X_{ij} for all i. If we are using linear regression or other means of prediction, this will yield a series of equations $f_i(X_{i(j)})$ for predicting the j^{th} values represented by X_{ij} . For the equations f_i that yield a good fit, we might consider any value $f_i(X_{i(j)})$ that differs significantly from X_{ij} as an outlier representing a possible error in the i^{th} data record.

29. The $f_i(X_{i(j)})$ are aggregates associated with an analysis. Any severe deviation of the $f_i(X_{i(j)})$ from X_{ij} could affect certain analyses significantly. We note that this general situation includes the existing editing situations as given by DeWaal (2003) or Draper and Winkler (1997). It corresponds roughly to the multivariate outlier detection situations of Kosinki (1999) and Rocke and Woodruff (1996). A prime consideration is that we may need only a few equations $f_i(X_{i(j)})$ to detect outliers and 'correct' the original data records. Other types of data errors that are not outliers and lie in the interior of distributions may be best detected by graphical methods (DesJardins, 1998) or inlier detection methods (Winkler, 1997).

30. For discrete data, subject-matter experts are likely to remain a good source of most or all of the edits. Additional edits might be determined by some exploratory data analysis. For instance with household data, it might be possible to look at the age distributions of pairs of individuals within a household. If, say, 0.1% of wives are 50 or more years older than their husbands, then a possible edit might be that an error occurs when the wife is 50 or more years older than her husband. If, say, 0.2% of children are older than their parents, then a possible edit might be that an error occurs if a child is older than the youngest parent.

IV. DISCUSSION

31. Although we have covered editing methods for detecting errors in values in data records, we have not described imputation methods. At the simplest level, any replacement value for a variable X_{ij} that causes the ith record to satisfy edits is acceptable. At a much more sophisticated level, we would need to impute (find replacement) values that also satisfy edits. For general data that does not cover edit restraints, Little and Rubin (1987) provide methods of imputation that can preserve the joint probabilities of variables that are needed for analysis. For discrete data, Winkler (2003) provides imputation methods that yield data that satisfies edit restraints and probabilistic restraints under the missing-at-random assumption.

32. For continuous data, edits may be determined by subject-matter specialists and elementary outlier-detection methods. For discrete data, edits are likely to be determined primarily by subject-matter specialists. Some exploratory data analysis may yield additional edits.

V. CONCLUDING REMARKS

33. All data analysis begins with a model. Most analyses of data depend on aggregates that can be computed from data. A loglinear analysis may depend on the joint distribution of a set of variables. A regression analysis may depend on how well a set of variables predicts another variable. In some instances, errors involving pairs of variables may be delineated and used in edit/imputation of the data. Many edits involving erroneous values of variables may be determined via subject-matter experts and by exploratory data analysis. If data are 'corrected' via valid edit/imputation procedures, then the data may be used for multiple sets of analyses.

References

Agrawal, R., and Srikant, R. (2000), Privacy Preserving Data Mining, *Proceedings of the ACM SIGMOD 2000*, 439-450.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

DesJardins, D. (1998), "A New Graphical Techniques for the Analysis of Census Data", *Statistics Canada Conference Proceedings*.

DesJardins, D., and Winkler, W. E. (2000), "Design on Inlier and Outlier Edits for Business Surveys," *Proceedings of the International Conference on Establishment Surveys, II*, 547-556.

De Waal, T. (2003), *Processing of Erroneous and Unsafe Data*, ERIM Research in Management: Rotterdam.

Draper, L., and Winkler, W.E. (1997), "Balancing and Ratio Editing with the new SPEER system," *American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods*, 570-575 (also available as Statistical Research Division Report rr97/05 at http://www.census.gov/srd/www/byyear.html).

Fellegi, I. P., and Holt, T. (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, to appear.

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*, New York: J. Wiley & Sons, 415-435.

Greenburg, B. and Petkunas, T. (1990), "Overview of the SPEER System," Statistical Research Division Report 90/15 at http://www.census.gov/srd/www/byyear.html.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Hidiroglou, M.A., and Berthelot, J.-M. (1986), "Statistical Editing and Imputation of Periodic Business Surveys," *Survey Methodology*, *12*, 73-83

Kosinki, A.S. (1999), "A Procedure for the Detection of Multivariate Outliers," *Computational Statistics and Data Analysis*, 29, 145-161.

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.

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 Luguna, Tenerife, Spain.
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 Luguna, Tenerife, Spain.
Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047-1061.

Van De Pol, F., and Bethlehem, J. (1997), "Data Editing Perspectives," *Statistical Journal of the United Nations ECE*, 14, 153-171.

Winkler, W. E. (1997), "Problems with Inliers," paper presented at the European Conference Statisticians, October 14-17, 1997, Prague, Czech Republic, http://www.unece.org/stats/documents/1997/10/data_editing/22.e.pdf

Winkler, W. E. (2003), "A Contingency Table Model for Imputing Data Satisfying Analytic Constraints," *American Statistical Association, Proc. Survey Research Methods Section*, CD-ROM, also research Report SRS 2003/07 at http:/www.census.gov/srd/www/byyear.html.
