Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

Discussants: Ton de Waal (Netherlands) and Maria Garcia (United States)

1. Introduction of topic

This topic covers new methods of data editing and imputation. The contributions could report on new methods and software, implementations in different situations, or organizational issues related to the development or implementation of new editing and imputation methods. Topics of interest include:

- the detection and treatment of systematic errors;
- the detection and treatment of outliers;
- development of new software applications;
- impact of new systems on organization efficiency;
- the use of graphical editing techniques;
- temporal editing and imputation;
- selective editing techniques;
- imputation under edit restrictions;
- outlier-robust imputation;
- variance estimation in the presence of imputation; and
- machine learning methods for editing and imputation.

Presentations of new methods should be linked with an evaluation of the operational feasibility of applying the ideas in a variety of data or organizational situations. Empirical comparison studies and evaluations are particularly welcome. These studies and evaluations could focus on aspects such as the processing time, the requirements for storage capacities, and expected impact of edit systems on data quality.

2. Summary of Papers

Invited Papers:

WP 30: Methods and Software for Editing and Imputation: Recent Advancements at ISTAT – *Marco Di Zio, Ugo Guarnera, Orietta Luzi, and Antonia Manzari, Italy*

Paper summary

The paper reports on the most recent research and implementation efforts at ISTAT on three separate edit and imputation problems:

1. *Editing-detecting unity measure errors*
The authors describe a new method for detecting a very common error in surveys, namely, reporting in the wrong units or unity measure error. The authors framed the problem as a classification problem in which records are classified by their reporting error patterns where each separate pattern represents a cluster. Error localizing wrong unity measure can then be setup as a cluster analysis problem. The paper describes a finite mixture model where each component represents an error pattern. In addition to using the model to identify records with erroneous unity measures, the parameters from the mixture models are also used for identifying outliers and detecting records that may have a potential high impact on the final estimates (selective editing).

2. *Imputation - Discrete and Continuous data*
ISTAT had developed a new software for imputation of discrete data based on using Bayesian networks to construct a model for computing the conditional probabilities of missing variables when the given variable depends on some other non-missing variables (non-missing parents in the network). The authors report they modified the algorithm by also including the non-missing children in the computation of the conditional probabilities. The missing value for the given variable is then imputed according to the conditional probabilities. The paper reports the model provides representations that do a better job of preserving the joint distribution of the variables.

For imputation of continuous data, ISTAT had developed new software, called QUIS. QUIS has three different imputation options: Regression imputation using the EM algorithm, multivariate predictive mean matching which also uses the EM algorithm, nearest neighbor donor imputation, and multiple imputation. QUIS has a user-friendly interactive windows interface and is written in SAS and SAS/IML.

3. *Editing Demographic Census Items-Enhancements to DIESIS*
The editing and imputation of demographic census items in DIESIS is based on the identification of passed records that could serve as potential donors for failed records that are similar to the donor records. DIESIS includes a new approach for selection of donor records based on partitioning the domain of passed records into smaller subsets with similar characteristics. The partitioning of the domain of passed records is done by solving an unsupervised clustering problem. Search for donor records is now done within the smaller clusters that are more similar to the failed record. The paper reports the new algorithm is computationally inexpensive making it suitable for very large census data sets. The algorithm is implemented in C++ and had been thoroughly tested for quality of donor selection and editing/imputation quality.

Points for discussion:

1. The finite mixture model makes a normality assumption, authors had done experiments in data that departs from the assumption, results still good but with extreme departure form normality method is expected to fail. Diagnostics from the model for detecting unity measure error can also be used for selective referrals of critical records.
2. Bayesian networks used only for imputation of categorical data, joint distributions better preserved, compared to hot deck?
3. Can the new donor search algorithm in DIESIS be used with other data besides census data? Will it still be useful when used with survey data as compared to census data (large so computational improvement is welcome)? Authors point the algorithm seems to be promising to impute demographic data for households with uncommon structures for which few suitable donors in general can be found.

WP 31: Smoothing Imputations for Categorical Data in the Linear Regression Paradigm – *Yves Thibaudeau and William Winkler, United States.*

Paper summary

The paper presents an alternative methodology for imputing missing data or inconsistent data that had been set to missing by the edit system. If a suitable number of donors is not available for hot-deck imputation then model-based methods such as log-linear models can be used. A linear regression model could be used instead of a log-linear model. The linear regression model has the advantage that the predictor and response variables are easily identified and the complex variance structures can easily be integrated. The authors present a method for filling in missing data based

on a linear regression model in which the conditional probabilities are estimated by estimating the conditional log-odds.

The authors consider a contingency table defined by four discrete variables and the process generating the table is described using an all-2nd order interactions log-linear model. The conditional probabilities are used to construct a set of log-odds. The log-odds are then expressed in terms of the conditional probabilities. The parameterization used for the estimation of the log-odds lead to framing the estimation of the conditional probabilities using a linear regression model. The model first performs a weighted least squares regression with the observed log-odds as the response variables and the parametric log-odds as the predictor variables. The estimates from the regression are then used to compute an estimate of the conditional probabilities. The paper then describes how the regression model can be extended to solving the problem of imputing missing data or replacing inconsistent data in edit failing records.

Points for discussion:

1. What are the advantages/disadvantages of using a regression model vs. using a log-linear model?
2. Approach seems to be a special case of the EM algorithm
3. What are the expected results/difficulties of applying this methodology to real census data?

WP 32: Using a Quadratic programming Approach to Solve Simultaneous Ratio and Balance Edit Problems – *K. Thompson, J. Fagan, B. Yarborough, and D. Hambric, United States.*

Paper Summary

This paper presents a methodology for solving ratio and balance edit problems based on minimizing the weighted distance between the reported and edited items in two separate surveys. Items in the Economic Census are first subjected to ratio edit tests and then items in balance edits are adjusted to ensure details add to totals. When using this two step procedure, items that are adjusted to satisfy balance edits may now fail the ratio edit tests. The situation is further complicated if data items must satisfy 2-dimensional balance constraints. If the balance and ratio edits are consider simultaneously then a quadratic program may be used to successfully find solutions that satisfy both the balance and ratio constraints. The approach described in the paper is a new use of existing ideas as the methodology can be compare to a least squares estimation for the data adjustment problem.

The methodology is a hybrid of ratio editing and quadratic programming. Items that must satisfy ratio restrictions are first edited to ensure the ratios are within the prescribed bounds, followed by solving a quadratic program to ensure both ratio and balance edits are satisfied, and control rounding to find integer solutions. The objective function is the sum of the weighted distance between the reported and imputed values and the constraints are the ratio, balance and non-negativity edits. The authors present results on two economic census applications from the Manufacture Sector and the Wholesale Trade sector. The paper reports major success using this approach for ensuring 2-dimensional balance in the example from the Manufacture sector. For the Wholesale Trade sector the initial solutions were not acceptable to the analysts. To obtain feasible solutions the authors report splitting the original quadratic program into two separate problems, attempt control of adjustment using a different weighting scheme, and removing some items from the objective function. Using this approach was a big improvement over 100% manual adjustment.

Points for discussion:

1. Using methodologists and subject matter experts in the design, retaining analyst preferences.
2. Success in replacing 100% manual review. New approach considers all constraints simultaneously which was not feasible with manual system.
3. Tested in two separate Census applications, use of available software (simplex algorithm)
4. When using a quadratic program the objective is to minimize the weighted distance between reported and edited record, which means all items could possibly be changed as opposed to minimizing the number of changes in the corrected record—not an error localization solution. Does not try to preserve distribution of details.

Supporting Papers:

WP 33: Data Editing and Logic – *Agnes Boskovitz, Rajeev Goré and Paul Wong, Australia*

Paper summary

The aim of this paper is to convert the error localisation problem to a corresponding logical problem. Logic is the study of deduction, and most methods of automated error localisation can be seen as a trade-off between search and deduction, where search means systematically testing potential corrections to the erroneous record, and deduction means finding some set D(E) of edits logically implied by the set E of pre-defined edits.
Logic gives two benefits. Firstly, it gives an alternative way of analysing the problem. Secondly, its collection of sophisticated automated tools could potentially be modified to use covering set correctibility for solving error localisation problems. The overall plan of the paper is to use logic to formalise the deduction component, as opposed to the search component, of automated error localisation.

In a 'pure deduction method', such as the Fellegi-Holt method, no search is needed at all. All of the pure deduction methods depend on the so-called 'covering set method'. When the covering set method finds exactly all of the error localisation solutions we say that the method of constructing D(E) is '(smallest weighted) covering set correctible'. In this paper "covering set correctible" is formalised in terms of logic.
This paper presents the beginnings of a theoretical logical framework. The main conclusions (for both categorical and arithmetic edits) are:

1. The generation of new edits can be seen as logical deduction;
2. The covering set method for error localisation is successful exactly when the deduction function has covering set correctibility, which is a strengthening of the logical constructs "refutation completeness" and "soundness";
3. The error localisation problem is a strengthening of the so-called satisfiability problem arising in mathematical logic;
4. The directional resolution method of solving the satisfiability problem depends on refutation completeness and soundness in just the same way as the covering set method of solving the error localisation problem depends on covering set correctibility.

Points for discussion:

1. Can operations research techniques be combined with techniques from mathematical logic?
2. Are SAT solvers faster (or more flexible) than operation research techniques (and traditional Fellegi-Holt techniques)?

WP 34: The Transition from GEIS to Banff – *Chris Mohl, Yves Deguirre, Robert Kozak, and Chantal Marquis, Canada*

Paper summary

This paper is reporting on the new generalized edit and imputation system developed at Statistics Canada. The new system, called Banff, has similar functionality as GEIS, allows the use of additional computing platforms and is more user-friendly. It is written in C and SAS/Toolkit for the SAS environment and can be run in a PC.

The paper starts with a brief history of GEIS and the reasons why a new generalized system written for the SAS environment was needed. Experience with GEIS was used as a starting point for building the new generalized system: the edit/imputation of GEIS was kept, however the different edit/imputation modules can now run independently of each other and is flexible enough to allow addition of future new options.

Highlights of the software:

- Collection of SAS procedures, SAS code generator driven by metadata, accept any input data format available in SAS, implemented in a modular approach with each procedures running independently producing outputs that can be used by any other component.
- Includes a library of C functions with the Banff algorithms.
- Linear programming interface.
- Error localization handles a larger number of variables than GEIS.
- Thorough testing and validation process done by methodologists and IT staff, each module tested independently, output compare to GEIS output, ensure software conforms to architecture standards.
- Software had been well received by users, users welcome a generalized E/I system in SAS, flexibility, and increased performance.
- Successful transition form GEIS to Banff at another statistical agency (ISTAT).
- No capability for handling discrete data, may be considered as a future option.

Points for discussion:

1. Redesign and implementation of an existing generalized edit/imputation system easily integrated in the complete survey processing.
2. The capability of running in the PC and using the common computing language for processing survey data has widened the use of the generalized E/I system at the agency. Is there a need at other statistical agencies of migrating to SAS processing?

WP 35: Edit and Imputation for the 2006 Canadian Census – *Michael Bankier, Canada*

Paper summary

The paper describes enhancements to the Canadian Census Edit and Imputation System (CANCEIS) in preparation for the 2006 Census. CANCEIS was successfully used in the 2001 Census to implement minimum change donor imputation for almost half of the census items, including all demographic variables. Donor imputation of the remaining census variables and deterministic imputation was done using separate software, called SPIDER. CANCEIS software enhancements in placed for the 2006 Census includes the capability of processing all census variables and performing deterministic imputation. This requires the inclusion of new modules

implementing methodologies previously used in SPIDER. The 2006 CANCEIS version will also include new tools such as a Decision Logic Table Editor, new distance measures for editing and imputation of continuous variables, the capability to process alphanumeric variables in addition to discrete and numeric variables, and simplified windows interface. Previous versions of the software had been used and validated at other Statistical agencies.

Points for discussion:

1. User friendly features, windows interface
2. Capability of porting/adapting modules that data experts had been using into the new software.
3. Collaboration amongst methodologists, programmers and subject matter experts.

WP 36: Integrated Modelling Approach to Imputation and Discussion on Imputation Variance – *Seppo Laaksonen, Finland*

Paper summary

During recent years the edit and imputation strategy at Statistics Finland has developed into an integrated modelling approach (IMAI). The strategy has been examined and further developed in two large research projects of the European Union: EUREDIT and DACSEIS.
The EUREDIT project concentrated on developing and comparing different methods for editing and imputation. EUREDIT tried to make imputations so that those preservation measures (concerning preservation of 'true' values) performed as well as possible. The DACSEIS project concentrated on variance estimation, including some experiments with imputed data.
This paper explains the principles of the IMAI approach. This approach is rather general and can be used both in the context of 'standard methods' and new methods. The approach also works with multiple imputation (MI). The IMAI approach does not see MI as any special imputation method, but as a potential tool for estimating the uncertainty of imputations (imputation variance). In the IMAI approach one first has to concentrate on finding the best possible (single) imputation technique, as done in EUREDIT, and –once satisfied with this – one has to try to provide a reliable uncertainty measure. For this latter step, MI can possibly provide a solution.
The IMAI approach is based on the following five steps: selection of training data and auxiliary variables, construction of imputation model, choice of criteria for imputation, the imputation task itself, and the calculation of the uncertainty for the estimates (including the imputation variance). There are several specifications for IMAI methods in Statistics Finland; most of them have been done using SAS. Standard software has not yet been developed. The main problem of the IMAI approach for a user is to decide the optimal options for each particular step of the strategy.

Points for discussion:

1. To what extent should we automate the edit and imputation process?
2. Can more guidelines for the IMAI approach be developed?
3. To what extent can we develop a systematic way of applying IMAI?
4. Is imputation variance an important issue at the moment, or should we (still) focus on imputation bias?

WP 37: Concepts, Materials, and IT modules for Data Editing of German Statistics – *Elmar Wein, Germany*

Paper summary

A joint work group of Destatis, the Federal Statistical Office (FSO) and the statistical offices of the Länder developed a new data editing concept. The FSO implemented the new concept in July 2004 via the introduction of materials, software, and IT modules. They support different activities related to (the planning of) data editing. Efficiency will be also promoted by a standardisation of data editing sub processes which is an important precondition for the reuse of available methods and respective IT tools. The new IT tools support the manipulation of metadata as well as the editing of data.

There are several IT tools for the planning of data editing, namely tools for the collection and judgement of relevant information, for the overall planning on the basis of predefined process chains, for specifying edit checks, and for detailed planning of data editing processes with process managers. The entire process chain is rather complex because it documents all possible data editing strategies. It begins with preparing activities (optional sub process) and terminates with the mandatory provision of (partly) plausible (meta) data for subsequent analysis/tabulation.
There are also several IT tools and materials for data editing. The IT tools for data editing are embedded in a concept which is based on the three standardised and modifiable sequences of data editing subprocesses. It is assumed that they cover a considerable majority (> 80%) of all existing data editing strategies. The processes are subdivided in typical activities and adequate methods are assigned to them. They form a collection of data editing methods, (graphical) analysis methods, coding methods and general data manipulation methods. The collection of methods shall represent a tool box. The FSO tested successfully a combination of a simple selective and macro editing method in 2004. It was decided to prioritise erroneous records of a stratum by a selective editing method and to prioritise strata (branches) by a macro editing method. Both methods were realised with SAS macros. Later it was decided to enhance the flexibility of the SAS macros. This led to the development of a macro template. A template is a macro that consists of the SAS macros mentioned above and may be supplemented by survey specific data steps. Finally, IT tools have been developed for automatic editing. For automatic corrections the FSO will use IVEWare.

Points for discussion:

1. Should we develop large software systems, or invest in developing several related modules?
2. Should we develop generic software tools, or software tools for particular applications?
3. To what extent should we aim to automate the statistical editing process and related processes?
4. How can we ensure that generic software tools are flexible enough?
5. Is the software tool fast enough for large/complex data sets?

WP 38: New Procedures for Editing and Imputation of Demographic Variables – *Gianpiero Bianchi, Antonia Manzari, Anna Pezone, Alessandra Reale, and Giorgio Saporito, Italy*

Paper summary

The paper reports on three new procedures for editing and imputation of demographic census items developed at ISTAT. The authors first described a methodology to reduce the low imputation quality and loss of information that may occur when connected groups of variables are handled in sequential edit/imputation steps. The methodology applies to demographic items and person items from the Italian population census. These items are processed in sequential edit/imputation steps, demographic items are handled first, followed by individual items. Since some demographic variables are connected to individual variables the loss of optimality could occur. The new method uses ideas from graph theory representing the questionnaire by a

connected graph. The graph is used to determine the variable with the most connections to other variables in a separate group. It then defines a new auxiliary variable and a subset of admissible values to control for the value of the variable responsible for the most connections between the subsets of variables.

In the paper the authors also describe a procedure for locating the household reference person when the edit checks indicate it's missing or more than one reference person exists in the household. The procedure is based on optimization techniques and had been carried out adapting the error localization algorithm implemented in DIESIS. The algorithm assigns the role of Person 1 to the person that minimizes the number of changes needed for the record to be consistent.

The paper also reports on the joint use of the "data driven" and "minimum change" approaches to treat erroneous responses for demographic variables. The first method finds a set of donors within the passed records (data driven). Then, determine the minimum number of fields to impute given these donors. The alternative choice is to first solve the error localization problem to determine the minimum number of fields to impute and then find the appropriate donors. The data driven methodology is selected as the default one with the option to turn to the algorithm implementing the minimum change method if the number of changes proposed by the data driven method was too large when compared to the number of changes proposed by the minimum change algorithm. The data driven approach is the preferred choice because it does a better job of preserving the population distributions.

Points for discussion:

1. Combination of various procedures for solving edit/imputation problems is a good strategy to correct records with complex structures/constraints.
2. Tests show that is not necessary to always the minimum change approach.

WP 39: Slice 1.5: A Software Framework for Automatic Edit and Imputation – *Ton de Waal, Netherlands*

Paper summary

This paper describes SLICE 1.5. SLICE is a software framework for automatic edit and imputation developed by Statistics Netherlands. This framework consists of several related modules. Recently version 1.5 of SLICE has been released for use at Statistics Netherlands. At Statistics Netherlands a combination of editing techniques is used to edit structural annual business surveys. One of those steps is automatic editing. For this step SLICE is used. SLICE 1.5 can handle a mix of categorical, continuous, and integer-valued data. To localise and correct erroneous fields SLICE uses edit rules. These edit rules check whether the data of each individual respondent are consistent. The Blaise parser of SLICE 1.5 can rewrite edit rules in Blaise format to edit rules in SLICE format that can be processed by the other modules of SLICE 1.5. Based on the edit rules the *Cherry Pie* module can localise the erroneous fields. The fields localised as being erroneous are set to missing. These fields can, if desired, be imputed by the imputation module of SLICE 1.5, together with the missing values in the observed data. While imputing one generally does not take edit rules into account. After imputation edit rules may hence still be violated. The *AdaptValues* module of SLICE 1.5 can adjust the imputed values so that all edit rules become satisfied. In this paper we give an overview of the edit and imputation process by means of SLICE 1.5. Subsequently, the paper discusses the Blaise parser, *Cherry Pie*, the imputation module, and the *AdaptValues* module.

Points for discussion:

1. Should we develop large software systems, or invest in developing several related modules?
2. Should we develop generic software tools, or software tools for particular applications?
3. To what extent should we aim to automate the statistical editing process and related processes?
4. How can we ensure that generic software tools are flexible enough?
5. Is the software tool fast enough for large/complex data sets?

WP 40: Imputation of Data Subject to Balance and Inequality Restrictions Using the Truncated Normal Distribution – *Caren Tempelman, Netherlands*

Paper summary

This paper uses the truncated (singular) normal distribution in order to obtain imputations that immediately satisfy both balance and inequality restrictions, while preserving the distribution of the data. First, the author derives the first order conditions for maximum likelihood estimation when the data are truncated non-singular normal. Since the solutions to the first order conditions are not available in closed form, an iterative procedure to obtain the maximum likelihood estimates using Fisher scoring is applied. Moreover, these first order conditions are difficult to compute because they involve multidimensional integrals that do not have closed forms or rapid numerical solutions. Therefore Monte Carlo integration is used in order to approximate these integrals. For this one needs to draw independent values from the integration region. The paper explains how this can be done. Next, the paper focuses on truncated (singular) normal distribution, and explains the differences with the non-singular case. Finally, the paper describes how maximum likelihood estimation can be carried out for the truncated (singular) normal distribution in the case of missing data. The resulting algorithm is a so-called MCEM algorithm. Giving the maximum likelihood estimates of the parameters, the truncated (singular) normal distribution can be used to impute missing data.

Points for discussion:

1. What is the quality of the imputations carried out by means of this approach?
2. What is the computing speed of this approach?
3. What are the largest (most complex) surveys that can be imputed by means of this approach?
4. Is this approach too complex?

WP 41: On the Imputation of Categorical Data Subject to Edit Restriction Using Loglinear Models – *Frank van den Eijkhof, Ton de Waal and Jeroen Pannekoek, Netherlands*

Paper summary

This paper describes a model based approach for imputation of categorical variables under edit constraints. The models considered belong to the flexible and widely used (for categorical data) class of loglinear models. The edit constraints are equivalent to the constraint that some value combinations must be zero (e.g. married=”yes” and age class= “0-10 year”). Such constraints are also known as structural zeros in the contingency table formed by all possible combinations of categories of all variables involved. The general approach is to estimate a constrained loglinear model for the contingency table with structural zeros and then use the model based estimated cell probabilities (for the non-structural zeros) to impute for missing values. Loglinear models usually do not consider more than 5 to 10 variables at a time, depending on the number of categories per

variable. When an imputation model is build for the simultaneous imputation of all categorical variables in a social survey, the number of variables can greatly exceed these numbers. Consequently, the dimensionality of the contingency tables involved and the number of cells in these tables can become very large. The paper describes the class of constrained loglinear models. Next, the paper explains how parameter estimation in moderately sized tables using the complete cases can be carried out. For parameter estimation for large tables this method can, however, not be applied because too many combinations have to be considered – the aforementioned dimensionality of the contingency tables involved. In order to overcome this problem the authors propose an approximation based on Monte Carlo methods. Giving the maximum likelihood estimates of the parameters, the constrained loglinear model can be used to impute missing data.

Points for discussion:

1. What is the quality of the imputations carried out by means of this approach?
2. What is the computing speed of this approach?
3. What are the largest (most complex) surveys that can be imputed by means of this approach?
4. Is this approach too complex?

WP 42: Evaluation of Score Functions for Selective Editing of Annual Structural Business Statistics – *Jeffrey Hoogland, Netherlands*

Paper summary

Since the statistical year 2000 Statistics Netherlands has a uniform statistical process for most Annual Structural Business Statistics (ASBS). The editing phase is an important part of this process, because filled in questionnaires for ASBS contain many influential errors. Statistics Netherlands uses a selective editing approach based on score functions to select records for manual editing of ASBS. Score functions that were used for ASBS 2000 were modified during the past few years. However, it was not clear whether these modifications actually improved error detection. This paper aims to evaluate some of the score functions used so far, using raw and edited data for nine publication cells for Retail trade and Transport. A questionnaire for an annual structural business statistic at Statistics Netherlands has four important blocks of variables, namely an employed persons block, a business profit block, a business costs block, and a business results block. The paper evaluates three types of score functions that monitor variables within a questionnaire block and two types of score functions that monitor key variables through ratios. The size of errors in records with sufficient scores hardly depends on the type of block score function or the type of ratio score function. The bias due to selective editing does depend on the choice for either a block score function or a ratio score function. The accuracy of reference values seems to play a decisive role.

Points for discussion:

1. Can selective editing be successfully applied to large/complex surveys?
2. Can current methods for selective editing be further developed?
3. Can a general theory for selective editing be developed?

WP 43: Automatic Editing System for the Case of Two Short-Term Business Surveys – *Rudi Seljak and Tomaž Špeh, Republic of Slovenia*

Paper summary

The Statistical Office of the Republic of Slovenia started to set up a system for automatic editing in the beginning of year 2004. They decided to start with the short-term business surveys. The reason for this decision was first of all the growing demand for quick results of these surveys and consequently the demand for quick and efficient flow of statistical process. The efficient method for data editing would naturally contribute a lot to the goal of shortening the period between the time when the data are captured and the time of dissemination of the results.
In the paper the application of two well-known editing methods on two short-terms business surveys, the monthly survey on turnover, new orders and value of stocks in industry and the monthly survey on wages, is presented. The two well-know editing methods are: the Hidiroglou-Berthelot method for detection of outliers and the Fellegi-Holt method for errors localisation and data imputation. The paper starts by giving some basic information on the surveys. It then presents the general approach which should be developed into the generalised system for short-term business data editing, and describes the application of this system for the case of the two above-mentioned surveys. Important tasks anticipated for the future are:

- to develop a computer program for derivation of implied edits out of the basic set of edits. For the case of the two discussed surveys this work has still been done partly manually.
- to determine up the most appropriate imputation method for the case of short-term business statistics;
- to develop a user-friendly computer environment for presentation of input matrices of the edit coefficients and the matrix on the correlates;
- to develop a system for efficient selective editing based on a score function approach.

Points for discussion:

1. Should we develop large software systems, or invest in developing several related modules?
2. Should we develop generic software tools, or software tools for particular applications?
3. To what extent should we aim to automate the statistical editing process and related processes?
4. How can we ensure that generic software tools are flexible enough?
5. Is the software tool fast enough for large/complex data sets?

WP 44: A Variable Neighbourhood Local Search Approach for the Continuous Data Editing Problem – *Juan-José Salazar-González and Jorge Riera-Ledesma, Spain*

Paper summary

The authors of this paper adopt the well-known Fellegi-Holt paradigm of minimum change to formulate automatic error localisation as a combinatorial optimization problem. They hence define the Error Localisation Problem (ELP) for a record as the problem to modify the fewest possible number of fields for the new, synthetic record to satisfy the set of edit rules. The objective of the ELP is commonly extended by considering the minimization of the weighted sum of the number of fields to be changed in order to satisfy the set of edits. In other words, a non-negative weight represents the confidence in the value of its related field. The paper focuses on continuous data, and consequently on the ELP for continuous data: the Continuous Data Editing Problem (CDEP). The combinatorial optimization problem that underlies in the CDEP has been shown to be NP-hard. Therefore the computation of the optimal solution for this problem may require an important amount of computational resources. The paper presents a new heuristic algorithm, based on a local search approach, to obtain a near-optimal solution for this combinatorial optimization problem. Some procedures of this local search approach make use of a Benders' decomposition of a linear programming model. The presented computational results

on randomly generated instances show that this approach solves instances up to 500 fields and 200 edits to near-optimality.

Points for discussion:

1. What is the quality of the solutions found by the proposed approach on realistic data?
2. How fast is the proposed approach on realistic data?
3. To what extent should we edit data automatically? Should we aim to edit all records automatically, or only the "easy" ones?

WP 45: An Editing Procedure for Low Pay Data in the Annual Survey of Hours and Earnings – *Salah Merad, Mike Hidiroglou and Fiona Crawford, UK*

Paper summary

The National Minimum Wage (NMW) was introduced in the UK in April 1999. Prior to 2004, the number of employees below the NMW was estimated using data from the New Earning Survey (NES) and the UK Labour Force Survey. Starting in 2004, only data from the newly redesigned NES, known as the Annual Survey of Hours and Earnings (ASHE) were used. Classification to the low pay domain (i.e. below the NMW) is based on hourly pay derived from total pay, hours worked, and the reporting pay period. This derived variable is subject to several sources of error, and can result in incorrectly declaring employees earning less than the NMW.
Selective editing is applied to the whole data set, but only detects a few of the erroneous records. This may result in biased estimates of the low pay counts. Consequently, it is necessary to have separate edit checks to obtain unbiased estimates, as it is important for decision making. These low pay edits use past and current values for a number of variables, as well as the current reported hourly rate. The latter is collected, as opposed to the derived hourly pay. A single edit check comparing the reported hourly rate to the derived hourly pay is obtained using a selective editing approach. Even though this reduces the number of records to follow up, this is not enough, as it exceeds available resources. The solution proposed in the paper is to adopt a hybrid editing procedure that follows up a smaller number of records in error, and yet maintains data quality. A sample of these failed records is validated, and their associated data are used to impute the remaining records. This is implemented in BANFF using two-stage donor imputation. The resulting imputed records are allocated in an unbiased way below and above the NMW. This holds globally, as well as at occupational group level, and within each gender group.
In this paper this hybrid editing procedure is presented in detail. The experience with its implementation using test data is also described, and its integration with the main imputation system is commented on.

Points for discussion:

1. Can selective editing be successfully applied to large/complex surveys?
2. Can current methods for selective editing be further developed?
3. Can a general theory for selective editing be developed?

WP 46: Implicit Linear Inequality Edits and Error Localization in the SPEER Edit System – *Maria Garcia, United States*

Paper summary

The paper describes updates to the US Census Bureau SPEER edit system. The SPEER software is a Fellegi-Holt system for editing economic data that satisfies ratio and balancing edits. The error localization algorithm in the previous version of SPEER generates a small subset of the

failing implied edits induced from failing ratio edits and balance equations for every edit failing record. The paper presents modifications implemented in this new version of SPEER that maintains its exceptional speed while doing a better job of error localization. The new version of SPEER uses Fourier-Motzkin elimination to generate a moderately large subset of the implied edits prior to error localization. The paper presents results of a feasibility study using Census data and a comparison study with the previous version of the software. The comparison study shows the new SPEER consistently corrects more records in two passes through the data with no significant gain in records corrected by running the data through the system a third time. The results also show the number of times a field is marked for deletion during error localization is reduced when using the implied edits generated prior to error localization. The study determined that previous heuristics in place to ensure the joint distribution of the variables is maintained were not needed when a large subset of the implied edits is available prior to error localization.

Points for discussion:

1. Error localization can be simplified if implied edits available prior to error localization. Effectively done at the Census Bureau for discrete data, now available for numeric data.
2. For continuous data it is not possible to generate implicit linear inequality edits prior to error localization—exponential growth of number of edits. Feasible because considering the particular type of survey data subjected to ratio and simple one level balance edits only.

WP 47: Improving an Edit and Imputation system for the United States Census of Agriculture - *Jeffrey Beranek and Robert McEwen, United States*

Paper summary

This paper is reporting on the procedures in place at the National Agriculture Statistics Service (NASS) for processing the US Census of Agriculture data. The paper gives a complete summary of the data capture, editing/imputation practices, and data analysis at NASS. The authors highlight that the 2002 Census of Agriculture was the first Census completely conducted by NASS after migration from the US Census Bureau. This represented a big challenge for the agency whose previous experience was mainly collecting data from sample surveys as opposed to censuses. For efficient management of the data processing NASS introduced new processes and re-engineered others. NASS decided to keep contract with the Census Bureau's National processing Center for the data capture process, used scanning and OCR technologies, upgraded the editing methodology to using decision logic tables (DLTs), used nearest neighbor imputation and interactive data analysis. The agency had various system problems including problems with the data bases, edit/imputation system design, and hardware configuration necessitating consulting with outside experts. The recommendations made by the outside consultants are integrated into the system for the 2007 Census.

For the 2002 census edit, NASS subject matter experts used Decision Logic Tables (DLTs) to encode if-then edits. The paper reports the development of a new "Authoring System" for DLTs in the form of SAS/SCL lists that eliminated the intermediate step of the analysts providing the edit specifications to the programmer, the programmer turning the DLT into code and then returning back to the analyst. The subject matter experts welcome the use of this authoring system with the capability of testing and implementing updates as needed. The DLTs are defined in smaller processing modules corresponding to logical subdivisions of the data. If the DLT determines there is an inconsistency then it tries to impute so that there are no edit failures. The imputation options include deterministic imputation, previous cycle data from some other NASS

data source, and nearest neighbor imputation. The system also has a utility for single record review. The paper also highlights the system improvements for the 2007 Census.

Points for discussion:

1. Importance of testing an entire system before production.
2. Importance of carefully designing the implementation of a series of major changes within a processing system.
3. Quality control measures (to be added for the 2007 Census)

WP 48: Improving Imputation: The Plan to Examine Count, Status, Vacancy and Item Imputation in the Decennial Census – *Arthur Cresce, Sally Obenski, and James Farber, United States.*

Paper Summary

The paper is reporting on the research and evaluation of alternative imputation methodologies at the US Census Bureau in preparation for the 2010 Census. The paper describes the methodologies being considered for imputing count items (size, occupancy, and status) and population characteristic items (sex, age, race, ...) The Census Bureau is considering matching census records to available administrative records for direct allocation of missing data. If records cannot be matched, then administrative records modeling will be implemented for the count items (housing unit status, occupancy, and size). Other imputation methodologies being considered are: Spatial modeling in which the person characteristics of neighboring units are used in the model; CANCEIS from Statistics Canada for sex, age, and relationship items; and the modified traditional hot-deck.

The paper also presents the strategies for evaluating the alternative imputation methodologies. The evaluation involves construction of separate truth decks for both count imputation and person characteristic imputation, inducing missing values in the true data using the observed missing values pattern from Census 2000, and studying the feasibility of using these methodologies in a production environment. The performance of all methods is to be evaluated by comparing the results of the imputation procedures against the reported (true) data. A separate truth deck consisting of vacant housing units will be used to evaluate the results of using the spatial modeling vs. the Census hot-deck for type of vacancy. The paper reports one methodology or a hybrid methodology will be chosen according to the results of the evaluation studies.

Points for discussion:

1. Agency's effort to determine whether the 2010 Census can use an imputation methodology that is feasible, cost effective, and produce more accurate results than the current procedures.
2. Assume a "true" truth deck can be build, measures for evaluating the effectiveness of the different procedures, can only compare methods that impute for same fields (not all methods are use for imputing any given field.
3. True data from 2000 used for evaluating procedures for 2010. Different data capture (electronic questionnaires, hand-held devices for follow-up)