WP. 34
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 (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

THE TRANSITION FROM GEIS TO BANFF

Supporting Paper

Submitted by Statistics Canada[1]

I. INTRODUCTION

1. Over the past thirty years Statistics Canada has developed a number of generalized survey software covering nearly all aspects of the survey process. Edit and imputation (E&I) systems have one of the longest histories of these products, dating back to the 1970s. In the mid 1980s, the Generalized Edit and Imputation System (GEIS) was created in order to meet Statistics Canada's economic survey needs with respect to edit and imputation. Although modifications and improvements had been made to GEIS on a regular basis, the significant change in the overall computing environment necessitated that a new system be built.

2. To address these requirements, starting in 2000, a new E&I system called Banff was developed. Unlike GEIS it can be run on a personal computer, uses SAS, the common computing tool employed at Statistics Canada, and does not require the user to perform steps in a specific order. In recent years, Banff has replaced GEIS as the most commonly used edit and imputation system at Statistics Canada, to a point where we expect that all GEIS users will have migrated to Banff within the next couple of years.

3. This paper describes the reasons for the development of Banff, the challenges faced and solutions employed when building it, as well as the experiences of users in making the transition from GEIS to Banff. In section II, a short history of GEIS from both the methodology and system points of view, along with some of its strengths and weaknesses is presented. The third section describes the requirements for this new E&I system. Section IV explains the characteristics of Banff, how it differs from GEIS from both the methodological and system standpoints, what associated tools have been built to interface with it and how it was tested. The reactions of users and the success of converting surveys from GEIS to Banff are presented in section V. In the last two sections we refer to potential future projects to further enhance the functionality of Banff as well as a summary on the overall project.

II. THE GENERALIZED EDIT AND IMPUTATION SYSTEM

4. Banff evolved from another Statistics Canada edit and imputation system called the Generalized Edit and Imputation System, or GEIS (Kovar, MacMillan and Whitridge, 1988, and GEIS Support Team, 1991). GEIS was developed for the E&I of numerical survey data, as part of the Generalized Survey Function Development Project of the 1980s. It has been used in production since the early 1990s.

[1] Prepared by Chris Mohl (Chris.Mohl@statcan.ca) Yves Deguire (Yves.Deguire@statcan.ca), Robert Kozak (Robert.Kozak@statcan.ca) and Chantal Marquis (Chantal.Marquis@statcan.ca)

5. GEIS has several restrictions. First, the incoming numerical data must be non-negative and continuous. However, it can also be used for edit and imputation of ordered qualitative data. Second, the edits used in GEIS must be expressed in linear form. That is to say, edits of the type "if-then-else" are not permitted. They can, however, be approximated. Finally, it is assumed that some preliminary editing has been done at the data capture stage, and that follow-up with respondents is complete.

II.1. Methodology

6. GEIS covers four main areas of editing and imputation: edit analysis, outlier detection, error localization, and imputation. Edit analysis consists of edit consistency analysis, the identification of redundant edits, deterministic edits and variables and hidden equalities, the generation of implied edits and extremal points of the feasible region described by the edits, and the production of summary statistics on the edit failure rates. Outlier detection uses a method proposed by Hidiroglou and Berthelot (1986) to perform univariate outlier detection, where values of selected variables are compared across records rather than comparing fields within each individual record. The error localization step, which determines the fields to be imputed, is carried out using Chernikova's algorithm (Chernikova, 1964 and 1965), according to the Rule of Minimum Change as proposed by Fellegi and Holt (1976). This means that the system minimizes the number of variables requiring imputation. The methods available for imputation are deterministic, nearest-neighbour donor, estimation via linear regression or other estimator functions, and unidimensional pro-rating.

II.2. System Information

7. At the time of the initial development of GEIS, the Oracle relational database management system was chosen as the foundation software because it was technically superior to many other options and portable across various computer platforms. The internal GEIS programs and routines were written in the C language. It was originally developed to be used on three computing platforms: the mainframe computer, UNIX, and in the MS-DOS environment on the PC. The PC version was soon discontinued however, mainly due to the insufficient processing power of the PCs of the early 1990s.

8. All of the GEIS E&I functions mentioned previously are built using a series of modules. These modules are all closely linked through a number of Oracle data tables. This makes it very difficult to use any of the modules independently, and implies a fixed order to the execution of the modules. The user must have an intimate knowledge of the design of the system in order to trick it so that, for example, donor imputation can be run without first running error localization. GEIS uses a "destructive" method of updating the data, overwriting the data that is already present in the tables before a module is executed. The user is encouraged to create backup copies of the tables before running a module.

9. Steps were later taken to provide independent modules in GEIS with the introduction of so-called "decoupled" modules for donor imputation, mass imputation, and pro-rating. That is, these modules could be used independently from the rest of the system through the use of a separate set of Oracle tables for the input and output data. Users wanted to be able to use these modules without having to go through all of the previous steps in GEIS.

II.3. Utilisation and Feedback

10. The methodology employed by GEIS is sound and widely accepted. When used properly, the system produces excellent, objective results. This led a significant number of surveys at Statistics Canada to use GEIS, usually around twenty at any given time at the height of its usage. Because GEIS could only be used on numeric or ordered qualitative data, almost all of these surveys were economic in nature.

11. Support for GEIS users is provided via a support team of both system and methodology personnel, written documentation and practice data files. With this help, surveys were able to implement

edit and imputation systems much faster and cheaper than if they had had to build them individually. However some drawbacks and limitations were observed.

12. As is always the case, a generalized system cannot meet the very specific or specialized needs of every survey. Even so, the full potential audience for GEIS was never fully realized due to the lack of flexibility and user-friendliness of the system, both real and perceived. In spite of encouragement from management to use it, many more surveys could have potentially used GEIS to meet their editing and imputation requirements had some of these issues been addressed.

13. In spite of efforts to decouple several key modules, the requirement to run the GEIS modules in a specific order discouraged some users, especially those who only wanted to use a sub-set of the functionality available in GEIS. The costs of working in an Oracle environment were significant as well. First, Oracle licenses are expensive and Statistics Canada did not have a corporate license for it. UNIX and mainframe platforms are also expensive. Furthermore, Oracle requires a steep learning curve on the part of the user, and the user must learn the SQL language in order to manipulate Oracle databases. As the computing power of the personal computer increased, more and more surveys moved to a Windows platform. Since GEIS did not operate on such a platform, it could not be used without migrating data back and forth across platforms.

14. There were some other limitations that impacted the widespread use of GEIS. For example, the number of variables that could be used in the error localization step was limited. Although there were many factors that influenced how many variables could be used, such as the nature and complexity of the edits as well as the platform on which GEIS was being run, it was generally found that, when the edits were of modest complexity, a maximum of about forty variables could be used. Beyond this number, the matrix of linear equations that Chernikova's algorithm produces grew too large for efficient processing. Also, the number of records processed by the donor imputation function had to be limited to control the processing times and costs. Finally, the lack of functionality for qualitative data meant that it was not useful for most social surveys and therefore was not adopted by this group.

15. Due to some or all of these limitations, certain surveys chose to develop their own custom-made system rather than use GEIS. While they met the functional needs of the survey, they were expensive to build and required extensive training for employees new to the system. At the same time, the ad-hoc development of a SAS-based prototype program which works on the PC for donor imputation, and which can also process negative and qualitative data took place. Some users, when given the choice of GEIS or this new program for donor imputation, selected the new program mainly due to its greater flexibility.

III. THE REQUIREMENTS FOR A NEW SYSTEM

16. It became apparent that a new edit and imputation system was required to address some of the major issues, such as those described above, which had dissuaded users from previously adopting GEIS. However it was also important that the new system have at least the same functionality and background methodology.

17. SAS had become a critical survey processing tool at Statistics Canada. In 1999, Statistics Canada adopted a corporate licence for SAS, allowing any employee within the organisation to use a large array of SAS products. Soon after, SAS was identified as a strategic tool for generalized systems especially for those at the heart of survey processing (sampling, edit and imputation, estimation). Therefore it became clear that a new E&I system had to be capable of running in a SAS environment, not only on a mainframe or UNIX platform, but also the PC.

18. The GEIS requirement that the steps be run in a certain order also had to be addressed. Too many users were either running unnecessary extra steps or not using GEIS because they were only interested in a subset of the functions available. The tricks that had been developed to work around this problem for

certain functions were too complex for many users. The new system needed to have decoupled functions so that each one could be run independently and in any order.

19. Another requirement was that it be flexible enough to allow additional functionality in the future. Although the original goal was to mimic the methods already in place in GEIS, over the longer term we wanted to be able to add in new functions or options to address some of the methods that were not available in GEIS.

20. Banff solves these problems. It is written to fully operate in the SAS environment and uses a component-based approach which allows users to run whichever functions they wish. The choice of SAS is obviously in line with the Statistics Canada strategy of using SAS in the development of generalized systems and also brings a rich environment highly suitable for running large batch edit and imputation jobs. The component-based approach also makes modification and improvements of individual functions much easier.

IV. BANFF

21. Having identified the requirements of the new E&I system, the programming of Banff started in 2000. In April of 2002, a beta version was released to allow users to become acquainted with it and feed back information to the developers. The first production version was released in January of 2004. In total, the equivalent of about fifteen person years of methodology and informatics work was required to reach this point.

IV.1. Methodology

22. The first version of Banff employed essentially the same methodology as that of GEIS, including the use of Chernikova's algorithm. The strategy was to first reproduce the methodological functionality of GEIS in a more user-friendly, more accessible, and more flexible SAS-based package, and then expand upon the methodology in subsequent releases. However, even without changes to the methodology itself, the way that it could be used differed from GEIS.

23. With Banff, users are able to execute E&I modules in the form of SAS procedures in any order that they wish, rather than being restricted by the fixed order in GEIS. This flexibility gives the user the ability to be more creative in the use of the available methods than they could with GEIS. For example, the user can now run only the prorating module to correct any simple inequalities between a total and the sum of its components if desired. Likewise, they can easily run donor imputation multiple times, using the imputation results of the previous executions to increase the chance of finding acceptable donor records in the current execution.

24. The ease of application of the Banff procedures in comparison to GEIS, and the fact that previous results are not automatically overwritten, also give the user the opportunity to very easily experiment with different parameter settings or various edit configurations. This improves the methodological framework and, in turn, the final results. The inflexibility made this a long and difficult (or in some cases impossible) task in GEIS.

25. In the latest release of Banff, some concrete methodological improvements have also been realized. Both the imputation by estimation and pro-rating procedures can now process negative data. Also, for imputation by estimation, the user can specify whether a random error term should be added to the imputed value to introduce some variability into the imputed data. This option is available whether the imputed value is calculated via linear regression or some other type of estimator (either pre-defined in Banff or custom-defined by the user). In GEIS, only linear regression estimators that were custom-defined by the user could have a random error term added to the imputed value. In the next release of Banff, the remaining procedures will also be able to process negative data.

26. A second method is being added to the prorating procedure. This method ensures that all eligible components of the sum are prorated in the same direction. That is, if the overall adjustment to the sum of the components is positive (negative), then all components eligible to be prorated will have a positive (negative) adjustment made to them. With the original prorating method, negative data could have negative adjustments made to them even though the overall adjustment to the sum was positive.

27. Finally, the outlier detection procedure has been modified to provide a more generalized form of the Hidiroglou-Berthelot method. The method can now be applied to the ratio of any two variables, rather than to only the ratio of the current value of a variable over the historical value of the same variable.

IV.2. System Information

28. Banff is a collection of nine SAS procedures designed to run on all available platforms at Statistics Canada: Windows 2000, UNIX (HP, SUN) and the mainframe. A recent experiment also determined that Banff could be run under Linux.

29. The software is written using the C language and the SAS/Toolkit. Within each Banff procedure the data is read in, transformed and output, ready to be used by any other component that can function in the SAS environment. As is the case for all SAS procedures, the Banff procedures are independent from each other and they are invoked by using the keyword PROC. Furthermore, they have a familiar look and feel to the procedures developed by the SAS Institute. For example, they use keywords such as DATA, VAR, ID, BY, OUT etc. The name reflects the functionality of the procedure. For example, the procedure to perform donor imputation is called PROC DONORIMPUTATION. The Banff procedures can accept any input data format that is available in SAS; data can come from SAS datasets, relational databases and PC files. Banff relies entirely on the capabilities of SAS to access data.

30. Internally, Banff is implemented in layers using a modular approach. Each layer provides a service or is a client to another layer. Central to this architecture is the Banff public application program interface (API), essentially a library of C functions that includes all of the Banff edit and imputation algorithms. Each function has a published interface and can be used by any application written using a third generation programming language. A Banff procedure uses a subset of these functions to perform a particular edit and imputation task. The modular architecture allows these algorithms to be re-used when adding new functionality. Another interesting layer in the Banff architecture is the linear programming interface. This interface allows Banff procedures to use external products to perform linear programming (LP) tasks. Currently, the Banff software can be dynamically linked to three known LP packages: GLPK, Xpress and CPLEX.

31. More efficient programming has also helped to address some of the problems related to capacity and run times in GEIS. The error localization function can now handle many more variables than was possible in GEIS. However a maximum number has not yet been established since, as in GEIS, the complexity of the edits plays a major factor in determining how efficiently PROC ERRORLOC can process the data. Also, because Banff runs in the SAS environment, which is optimized for batch processing, large jobs that involve many Banff procedures are run very efficiently.

IV.3. Associated tools

32. The Banff software can be used in a number of ways to perform the edit and imputation tasks of a survey. The first obvious way is to write a SAS program that calls the different Banff procedures via PROC statements in the desired order and deals with all of the input and output data requirements. This approach works well when performing testing and certification of the software since experienced Banff developers and users are involved. In these cases the programs need to test very specific aspects of the system and often need to be modified many times to fully cover all possible scenarios. Parameters within the SAS program can be quickly changed to do this verification.

33. On the other hand, production environments need a more stable and economical way of managing their SAS programs. Making use of programming resources to modify production systems frequently can turn into a costly experience, especially if there is a repeated turnover of staff or if changes need to be made in many parts of the program. To address this, the Banff team created what is now known as the "Banff processor". It is a SAS code generator that is metadata driven. This means that it is driven by information stored in tables. The information describes the edit and imputation strategy that is to be followed for a specific survey or questionnaire. The Banff processor recognizes keywords and generates calls to the Banff procedures. The tables where the information resides are owned and maintained by the project team responsible for the survey. It is currently the most economical and flexible approach to using Banff (in a production environment) when the edit and imputation strategy is complex.

34. Some surveys have very simple needs and are not recurrent. The development of metadata tables containing all the information that implements their edit and imputation strategy may represent more work than writing a SAS program. However in some cases, project team members may not have the knowledge or desire to write the SAS code required to call the necessary Banff procedures. For this reason, the Banff team decided to build visual interfaces to the Banff procedures. Those visual interfaces are available through SAS Enterprise Guide and are called wizards. The wizards offer a point and click environment, where values and statement variable names can be chosen from a list and dragged and dropped in the appropriate areas. SAS code is then generated behind the scenes. This SAS code can be submitted directly or saved and used later. This functionality will be available in the next release of Banff.

35. Banff being a replacement for GEIS, it is natural for GEIS users to migrate their application to this new software platform. However previous work done in GEIS does not need to be lost. The user who wants to minimize the impact of moving from GEIS to Banff can simply leave the GEIS data tables in Oracle. With the use of SAS/ACCESS for Oracle, the Banff procedures are capable of processing data that resides in GEIS Oracle tables. Another option is to import the GEIS Oracle tables into SAS as SAS datasets. The processing will then take place entirely in SAS.

IV.4. Testing

36. The mandate to redesign GEIS meant that the new product would contain at least all of the methodology of the old system, with possible enhancements, new methods and improvements to the performance. The testing of Banff was done by two different groups: one composed of IT specialists and another composed of methodologists.

37. The group of IT specialists was responsible for developing the software and used a software validation process. This process took place during the development phase in order to detect bugs as early as possible. The other goal of the software validation process was to ensure that the software was written to conform to the programming and architecture standards. Two approaches were taken: the white-box approach and the black-box approach.

38. The white-box approach allowed the team to review the design of the different Banff modules as well as to inspect the source code with the idea of finding defects as soon as possible. The whole IT group was involved in reviewing the modules. A side benefit was the learning opportunity for every member of the group.

39. The black-box approach is more traditional since it tests the behavior of modules irrespective of their implementation. The test material that was used consisted of the data coming from the Oracle data tables used for the testing of GEIS. Since Banff is implemented in layers using a modular approach, the first part of the testing was also modular. Each major algorithm was tested separately using appropriate input data. Each output was compared with the corresponding GEIS output or with hand-written expected results. Once modular testing was considered satisfactory, components were built. Each

component was an executable program that performed one of the nine functions of edit or imputation available in Banff. Integrated tests of these components were performed and the results were once again compared with the ones generated by GEIS or with our own custom built data. When integrated testing proved that all the algorithms had been properly redesigned, the SAS procedures were built. This included verifying that the SAS procedures behaved as any other SAS procedure would in the SAS environment. Testing here involved verifying that input and output datasets were created properly, that parameters were recognized and interpreted correctly, that errors were identified and output to the SAS log etc., and of course that results compared with the GEIS results.

40. Once the IT group was done with its testing, the Banff procedures were turned over to the methodology team for certification. The principle behind the certification is that before going into production the software should undergo a series of tests to ensure that the software meets the documented specifications on all supported platforms. These tests are more valid if performed by a group not involved in the software development, but who understands the methods being used. These testers have no pre-conceived ideas about the software that would limit the extent of their testing. Any new version of Banff always undergoes certification on each of the three platforms that it is designed to run on.

V. UTILISATION AND FEEDBACK

41. Reactions to Banff have been favourable. Since it is implemented for the SAS environment, it has been seen as an easy tool to learn and use. Current GEIS users have seen advantages in moving to Banff because of the flexibility it offers and because of its increased performance. It also gives users the opportunity to move to the Windows/PC platform on which more and more surveys are being processed. The fact that it comes with a SAS code generator driven by metadata has also helped to convince a number of users to convert to this software for their edit and imputation needs. Nearly all of the former GEIS users have migrated to Banff and we expect the rest to make the move in the near future. The Italian Statistical Bureau (ISTAT) made a successful transition from GEIS to Banff. Other international statistical agencies have either purchased, or are in the process of examining the possibility of adopting Banff for their E&I needs.

42. As with GEIS, user support comes in several forms. Informatics and methodology staffs are available to assist users in either making the transition from GEIS to Banff or starting up new surveys in Banff. Detailed descriptions of the methodology behind each of the SAS functions and documentation on how to code each function have been created (Banff Support Team, 2003, 2004). A tutorial for hands on learning about the functionality of Banff is currently being written to help those people less comfortable in writing SAS code.

43. One persistent negative comment concerns the lack of qualitative data functionality. Social survey project teams continue to be unable to use Banff since their data is categorical rather than numeric. These teams are forced to look elsewhere to fulfill their edit and imputation needs.

VI. POTENTIAL FUTURE WORK

44. Banff has established itself as the primary edit and imputation software for business surveys at Statistics Canada. Its capacity to perform numerous functions for quantitative and ordered qualitative data makes it ideal for such purposes. However there are several areas which could use further improvement, especially if Banff is to become a tool to be used for social surveys as well.

45. As stated earlier, one of the major limitations of Banff is that it is unable to process and impute unordered qualitative data. While this is not a major drawback for business surveys where the majority of responses are quantitative in nature, much of the data collected by social surveys is categorical. In order for Banff to expand its potential user base, the ability for it to process, edit and impute this kind of data needs to be added.

46. Specifying edits in the format required by Banff can be difficult for some equations. Non-linear equations cannot directly be input into Banff although they can be approximated in some cases by linear edits. Statistics Canada has developed a decision table software package called Logiplus (Systems Development Division, 2000) which allows edits to be expressed via decision tables and converted into SAS code. This program could be interfaced with Banff, allowing the user to express their edits via either traditional linear equations and/or the Logiplus interface. This functionality will be especially important if steps are taken to implement qualitative data processing.

VII. SUMMARY

47. The transition from GEIS to Banff required a significant amount of planning, development, testing and training, however the end result was a product that is more relevant to the current survey environment in Statistics Canada. The capability of running it on a personal computer has widened the use of the generalized edit and imputation tool and opened up a new potential audience. Its use of SAS and the associated SAS wizards, as well as the independence of the procedures make it more user-friendly to new users allowing them to understand and implement it more easily.

48. While Banff has proven itself to be valuable for economic surveys, there is still room for improvements to make it more applicable for social surveys. Additions such as the inclusion of qualitative data processing capabilities and improved edit writing capabilities will make this goal more feasible. As with all software, the developers will have to be constantly aware of new developments in the field of edit and imputation in order to determine what enhancements would be beneficial to the future of the product.

References

Banff Support Team (2003). Functional Description of the Banff System for Edit and Imputation System. Statistics Canada, Quality Assurance and Generalized Systems Section Technical Report.

Banff Support Team (2004). Banff v1.02 User Guide. Statistics Canada, Quality Assurance and Generalized Systems Section Technical Report.

Chernikova, N.V. (1964). Algorithm for finding a general formula for the nonnegative solutions of a system of linear equations. U.S.S.R. Computational Mathematics and Mathematical Physics 4, 151-158.

Chernikova, N.V. (1965). Algorithm for finding a general formula for the nonnegative solution of a system of linear inequalities. U.S.S.R. Computational Mathematics and Mathematical Physics 5, 228-233.

Fellegi, I.P., and Holt D. (1976). A systematic approach to automatic edit and imputation. Journal of the American Statistical Association 71, 17-35.

GEIS Support Team (1991, revised October 2000). Functional Description of the Generalized Edit and Imputation System (GEIS). Statistics Canada, Research and Generalized Systems Subdivision Technical Report.

Hidiroglou, M.A. and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys. Survey Methodology 12, 73-83.

Kovar, J.G., MacMillan, J. and Whitridge, P. (1988). Overview and strategy for the Generalized Edit and Imputation System. (Updated February 1991). Statistics Canada, Methodology Branch Working Paper No. BSMD-88-007E/F.

Systems Development Division. (2001). Logiplus User's Guide. Statistics Canada internal report.
