1. The Canadian Census is conducted every five years since 1956 and is Canada's largest and most comprehensive data source. The Census of Population collects demographic and linguistic information on every man, woman and child living in Canada. It is the main source of comprehensive statistical data, offered in a standardized format, at the sub-provincial level on Canada's population. It provides nationally comparable data that can be cross-classified to show details. For the 2011 Census, the short form was delivered to 100% of the dwellings and the voluntary long form (National Household Survey) was delivered to 30% of the dwellings.

2. The data collected from the Census and NHS were processed through editing and imputation (E&I) using CANCEIS, the CANadian Census Edit and Imputation System, originally developed in 1996.

3. This paper provides a brief overview of CANCEIS as well as recent changes that were implemented for 2011. In addition, the lessons learned from the 2011 production are presented, along with the current options that are being considered for 2016.
II. CANEIS Overview

4. CANEIS offers two components: a derive engine and a donor engine. The derive engine performs deterministic imputation and creates new variables, and the donor engine performs donor imputation. Due to its computational advantages, CANEIS can process very large data files, i.e., a great number of records and variables, and allows users to specify numerous edits. CANEIS can work with categorical, numerical and alphanumerical variables simultaneously. Around these two components, there are four types of input files provided by users and several output files produced by the system, as shown in Figure 1 below.

**Figure 1: CANEIS Components**

Below, the input and output files of CANEIS are first introduced followed by a description of the system components.

A. Input files

5. The input data file contains units to be processed through E&I. A unit could be a single record or it could consist of several records (subunits), such as household members within a household. Note that CANEIS can only process together units containing identical numbers of subunits.

6. In order to increase the efficiency of donor searching, similar units will appear on the file as closely together as possible based on subject matter expertise. For example, in the demography process of the Canadian Census, households are arranged according to their geographical proximity. This is based on the long-standing assumption that similar families tend to live in the same neighbourhood. However, in other situations, a different sort order may be preferable. For example, in the income process of the 2011 NHS, units were ordered by their total income when imputing for the income components. In turn, the subunits can be ordered within a unit according to chosen rules, and the evaluation of the similarity of two units is accomplished by evaluating the similarity of the corresponding subunits.

---

1 Version 5.1 of CANEIS is presented here (see Liu, Crowe, Alavi (2011)). Once version 5.2 is available, a new user guide will be provided and will describe the new features such as changes in contents, parameters, filenames, etc.

2 Although CANEIS does not do direct historical or regression imputation, it can handle workarounds, such as using variables (historical or from a regression model) that are correlated to the variables of interest as auxiliary matching variables in finding a donor.
7. The data dictionary provides information to be used in the module, such as the variable names and types, and all possible and valid values for variables (the latter applying to donor modules only). Users can construct a data dictionary with either text files or an Excel workbook. To facilitate the writing of edit rules, CANCEIS works with user-defined text labels for categorical variables instead of the associated numeric codes, e.g., “MALE” and “FEMALE” rather than “1” and “2”.

8. The system parameters play an important role in the flexibility, adaptability and efficiency of CANCEIS by providing users full control over the vast array of choices available. To name a few, users can have control on editing, the donor search, reordering of subunits, and E&I reports.

9. In CANCEIS, Decision Logic Tables (DLTs) are used to specify edit rules for both derive and donor modules. The edit rules defined in CANCEIS are, in general, conflict edit rules as the system is looking for anomalies and inconsistencies in the data. In a DLT, the user specifies conditions called propositions that link to one or more rules. If all the conditions entering a rule are true for the unit, then the rule is said to be matched. Derive DLTs and donor DLTs are used in different ways, however. In derive DLTs, the user specifies a special set of instructions called conditional actions that determine precisely how to handle the conditions. Derive DLTs also allow a set of general instructions called common actions that get applied to all units, which is useful in the creation of new variables. In donor DLTs, there are no actions. The matched rule indicates that the unit has an issue that will be corrected with donor imputation.

Figures 2 and 3 are examples of DLTs used in derive and donor modules, respectively.

**Figure 2: Derive DLT**

<table>
<thead>
<tr>
<th>Propositions</th>
<th>Edit Rule 1</th>
<th>Edit Rule 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE(#1) &lt; 15</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>MARST(#1) = CLASS(EVER_MARRIED)</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>INCOME(#1) &gt; 0</td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>MARST(#1) = SINGLE</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>INCOME(#1) = 0</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

**Figure 3: Donor DLT**

<table>
<thead>
<tr>
<th>Propositions</th>
<th>Edit Rule 1</th>
<th>Edit Rule 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE(#1) &lt; 15</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>MARST(#1) = CLASS(EVER_MARRIED)</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>INCOME(#1) &gt; 0</td>
<td></td>
<td>Y</td>
</tr>
</tbody>
</table>

10. In the above two figures, the propositions are listed in the first column. For edit rule 1 (the second column) the Y's identify two conditions: a person is less than 15 years old and has a MARST (i.e., marital status) of EVER_MARRIED (which is defined as being either married, divorced or widowed). A unit satisfying the two conditions of edit rule 1 are said to match the edit rule. The X in edit rule 1 of the derive DLT prescribes the solution (action) of deterministically changing the marital status to “single”.

3
On the other hand, the donor DLT would prescribe that a donor must be found so that either AGE or MARST, or both variables, will be imputed in order to resolve the conflict. Usually both AGE and MARST will not be imputed together since that would violate the minimum change principle.

11. In these DLTs, we can see two further features of CANCEIS. First is the expression CLASS(EVER_MARRIED) which refers to a group of values of the variable MARST, defined in the data dictionary by the users. If this feature is not used then it would be necessary to have separate edit rules for each of these values. A second feature pertains to writing common edits applicable to all subunits within a unit, only once, instead of repeating a rule for all subunits. To do so, the user can add #1 in brackets next to the variable in the edit rule, and the system will know to repeat this rule for all subunits within the unit.

B. Output files

12. At the end of an E&I module, CANCEIS creates a file of imputed data as well as some reports. In particular, as an optional output file, the audit trail from a donor module provides a very useful step-by-step report showing how the best donor for a given unit was found. Other output files provide imputation statistics such as the unit failure rate and how many times a given donor was used. The user controls the output files and the extent of their details with system parameters.

C. Derive Engine

13. The derive engine has the capacity to either create new variables or perform deterministic imputation to correct systematic errors based upon subject matter experience and knowledge. For this purpose, CANCEIS provides some of the same functionality as many computing languages, such as functions (e.g., assigning random values, finding maximum value), “DO loops” and “calling sub-routines”.

D. Donor Engine

14. The donor imputation component of CANCEIS is based on the Nearest Neighbour Imputation Methodology (NIM) (Bankier, 1999 and Bankier, 2011). As an alternative to the well known Fellegi-Holt methodology (Fellegi and Holt, 1976), NIM finds potential donors first and then decides on the minimum number of variables to be imputed based on each potential donor. The reversal of the imputation steps fulfills the data-driven approach and gives NIM significant computational advantages, while still meeting the objectives of imputing the fewest variables possible and preserving subpopulation distributions as much as possible.

15. Editing is conducted and controlled through the specification of validity classes (acceptable values for variables) and conflict edits defined within the DLTs. Each unit is evaluated to identify invalid values (values not in the validity class) and inconsistencies (conflict edits matched). Those units that have no invalid values and no inconsistencies are said to pass the edits and are referred to as passed units. The units that have at least one invalid value or inconsistency are said to fail the edits and are referred to as failed units. Conceptually, CANCEIS processes failed units successively in the order of their appearance, but in practice, CANCEIS is capable of independently processing several failed units simultaneously.
16. While the user can specify the edits in many small decision logic tables, CANCEIS creates a unified version of the edits by combining all the individual edits and deleting the redundancies. When imputing a particular record, it is often possible to drop many edits because the potential donor is identified first. For example, if there are no grandparents in either the failed household or the potential donor household being examined, CANCEIS will drop (i.e., ignore) all the edits referring to grandparents because they do not apply. Customizing the set of edits to each failed-donor pair decreases the processing time substantially.

17. For each failed unit, many passed units are considered and the nearest neighbours are identified. The nearest neighbours are defined across multiple dimensions by comparing a given set of matching variables. Sorting the data appropriately beforehand on key variables allows the closest nearest neighbours to be found in the least time, within the donor pool defined. Identification of the nearest neighbours is based on the similarity of the values between the failed and passed units for each matching variable and the weight (relative importance) of each matching variable. These concepts are made quantitative through the total distance measure $D_{fp}$ between the failed and a passed unit:

$$D_{fp} = \sum_i w_i D_i(V_f, V_p)$$

where $V_f$ and $V_p$ are the values of the $i^{th}$ matching variable for the failed and passed unit, respectively, $w_i$ is the weight assigned to the $i^{th}$ matching variable in the data dictionary and $D_i$ is the distance function selected for the matching variable that evaluates the similarity of $V_f$ and $V_p$. The weight assigned to a matching variable is usually reflective of its value as a predictor for the variables to be imputed. The choice of different weights for different variables is primarily based on subject matter expertise.

18. CANCEIS currently has ten distance functions $D_i$ (see Bankier 2011) which enable the system to process all types of variables. For example, the distance function #1 can be used for discrete, coded and alphanumeric variables. Using this distance function, if the response of the failed unit is the same as the response of the potential donor, then the distance is set to 0; otherwise, it will be set to 1. Another example is distance function #3 which is used for discrete variables and allows for a range of acceptable passed values around the failed record value, and all values in the range will each be assigned a distance between 0 and 1 depending how far they are from the failed value. That being said, each of the distance functions have user-defined parameters to allow even more flexibility. The distance measure $D_i$ in Equation (1) allows all types of variables to be brought together on the same scale, as the values of $D_i$ are between 0 (values “essentially equal”) and 1 (values “totally dissimilar”). This is why CANCEIS can treat various types of variables simultaneously. For a given failed unit, a list of the closest nearest neighbours is retained for further evaluation.

19. For a given donor module, each variable appearing in the input file is defined by the user to either be imputable or nonimputable. Both imputable and nonimputable variables could be used in the module as matching variables and could be used in the DLT propositions. During the imputation procedure, imputable variables that have invalid values will always be imputed immediately since this is necessary for any successful imputation action (IA), which specifies which values will be taken from the potential donor. In the simplest situation, the failed unit would then pass all edits and a potential IA is found. Otherwise, if inconsistencies are still present, then for each remaining imputable variable (for which $V_f \neq V_p$) the choice remains whether to impute or not. In order to facilitate an efficient search for potential IAs, a binary tree is employed wherein each node yields two branches representing the decision to impute or not impute for a particular variable. It is possible for a given nearest neighbour to generate more than one potential IA or no IA at all. For example, IAs consisting of imputing either age or
marital status might resolve the conflict in the first edit of the Figure 3 donor DLT. Each nearest neighbour from the list of best nearest neighbours is in turn evaluated, by ascending $D_{fp}$ for potential IAs.

20. For the purpose of deciding the best potential IAs, CANCEIS uses the quantitative measurement tool $D_{fpa}$ defined as

\[ D_{fpa} = \alpha D_{fa} + (1 - \alpha) D_{ap} \]  

(2)

where ‘a’ represents the unit after imputation, $D_{fa}$ and $D_{ap}$ are defined as in Equation (1), and $\alpha$ is a user-defined system parameter in the range (0.5, 1]. In general, we expect that the passed units most similar to the failed unit, implying lowest $D_{fp}$, are the most likely to yield the best or minimum-change IA. The $D_{fa}$ value measures the similarity of the imputed unit to the failed unit, thus measuring the minimum change aspect of the IA. On the other hand, the $D_{ap}$ value measures the similarity of the imputed unit to the passed unit, thus measuring the plausibility aspect of the IA because the potential donor has entirely real data. Note that to enforce the minimum change principle, the parameter $\alpha$ must be greater than 0.5. (More details in Bankier, 2011.)

21. CANCEIS keeps a running list of the best potential IAs (i.e., those with the lowest $D_{fpa}$) for a failed unit as the nearest neighbours are examined. The user specifies the maximum number of IAs to retain on the list, and how good an IA must be to bump another IA from the list. Once all nearest neighbours have been checked to identify all potential IAs, an IA is randomly chosen from the list of the best potential IAs as the actual imputation action for the failed unit.

E. Further Control and Efficiencies

22. The user has fine control over most aspects of the E&I process through various parameters. These aspects include staged donor searching, outlier detection and control, reordering of subunits, using failed units as donors, and the ability to customize the weights and distance functions for different types of failed units, to name a few. More details on donor searching are given below.

23. Donor searching is usually done in stages for two reasons. First, it may not be necessary or practical to evaluate all passed units for potential IAs. Second, it is often true that the best donor units are physically close to the failed unit, especially when units in the data file have been appropriately ordered in advance, with respect to desired characteristics. Through system parameters, users have control over how many units are considered as potential donors in each stage, and the maximum number of stages that can be done. Before going on to the next stage, CANCEIS will evaluate whether there has been significant improvement in the quality of potential IAs in the previous stage. If not, no further stages will be processed. The user can also specify an absolute minimum standard for the donor used to generate the IA so that passed units with too high a $D_{fp}$ relative to the failed unit will not even be considered. This will save time as the system will not be analyzing unacceptable donors for IAs, while ensuring that the final chosen IA is of a minimum standard. Based on the best IAs found so far, CANCEIS may conclude that there is no need to search further. These mechanisms increase efficiency substantially.
F. Major improvements for the 2011 Census

24. For 2011, CANCEIS was rewritten in C# (C sharp) computer language within the .NET environment to improve the efficiency of the CPU, allowing for multi-threading processing. Also, CANCEIS was made more user friendly, now accepting input data dictionary files in EXCEL format, and generating HTML format output files, as opposed to just dealing with input and output files in .txt format.

25. Prior to 2011, the ethnocultural (EC) variables were divided into five distinct processes, namely place of birth of parents, immigration, aboriginal, ethnic origin and visible minorities, and the edit and imputation was done separately, one process at a time. Anomalies were sometimes caused by imputation, given that up to eight different donors could be used together to impute one record. For 2011, some functionality has been added to CANCEIS specifically to edit and impute the variables from all five of the old EC processes at the same time. Users are now able to specify different parameter values for donor imputation by type of unit. This improvement overcomes the problem of using a single set of parameter values for vastly different types of units. Moreover, an optimal distance parameter was added to the algorithm that determines the stopping criteria when searching for donors. Although this latter option was added to improve CANCEIS in general, it turned out to be particularly crucial for the EC process.

III. Lessons learned with the 2011 production

26. From the 2011 production, we have learned that:
   a. The CANCEIS v5.1 offered more stability than versions used from previous censuses after being redeveloped in C#/.Net. CANCEIS was able to handle larger amounts of data with adequate performance, especially for the combined EC process where more variables (nearly 30) and more edits were treated together.
   b. Combining all five EC processes into one not only helped to reduce processing time, but also helped to improve data coherence between EC variables due to the use of a single donor to impute the variables for all five processes at once, rather than using up to eight different donors when done separately.
   c. Donor imputation requires a lot of processing power, so we plan to continue research and development to tune CANCEIS to even further improve performance.
   d. While the E&I methodology was adjusted in production to react to new patterns observed in the item non-response, CANCEIS offered flexibility and options to do so in a timely manner.
   e. Increasing the content and the level of detail of the LOG output file proved to be very useful and facilitated the work required.
   f. Being able to edit the data dictionary in an Excel workbook was also more user-friendly.
IV. Options currently considered for 2016

27. For 2016, discussions will take place between methodologists and subject-matter experts to determine if we should group more processes together, such as the home language, mother tongue, and official language processes. The multidisciplinary team will also assess if we could reduce the number of imputation steps done by process, by combining more imputation modules together, making the E&I strategy more efficient. We will also evaluate if some processes could be run in parallel, rather than sequentially, in order to reduce the overall processing time.

28. We plan to improve CANCEIS on an ongoing basis based on user needs and resources available. The CANCEIS development team has already started to review the current set of default values and default parameters implemented in CANCEIS to determine if these should be updated, to reflect the most common and “intuitive” values, based on our current experience and feedback from users.

29. CANCEIS v5.2 should be officially available sometime in 2014 which will reflect the above changes, and also include the following improvements: the naming convention of input and output files will be standardized and the content of each file will be revisited, making CANCEIS easier to manage, and with contents more intuitive. Complete documentation will also be made available to accompany the new system. In addition, to help current users migrate their metadata into the new v5.2 environment, a conversion tool, named CANVERT, has been developed and will be added to the installation package.

V. Conclusion

30. CANCEIS\textsuperscript{3} performed very well for the 2011 production and provided solutions to all challenges encountered. In summary, CANCEIS can be beneficial for E&I users who require:

- a system that can perform deterministic and donor imputation as well as derive new variables,
- the ability to process a great number categorical, numerical, and alphanumerical variables simultaneously,
- the ability to easily define large numbers of edits,
- the capability of processing large data files quickly and efficiently,
- the flexibility to have fine control over all aspects of the process through simple user-defined parameters, and
- software which can be used immediately on most computing platforms without the need of complex installation of custom programs.

\textsuperscript{3} To obtain the most recent official version of the CANCEIS software, the CANCEIS executable (.exe) and documentation are available at no cost for any agency interested in using it, once they have signed a license agreement.
31. CANEIS is constantly evolving and could potentially be used by a more and diverse group of users, in contexts beyond censuses and other social surveys, who wish to apply deterministic and nearest neighbour donor imputation to their survey data.

VI. References


