SSC Annual Meeting, June 2011
Proceedings of the Survey Methods Section

STATISTICS CANADA'S SURVEY ON COMMERCIAL AND INSTITUTIONAL ENERGY USE: AN APPLICATION OF INDIRECT SAMPLING

Steve Matthews[1] and Tyler Kirkland[2]

ABSTRACT

The Survey on Commercial and Institutional Energy Consumption (SCIEU) is conducted by Statistics Canada on behalf of Natural Resources Canada and Environment Canada on an annual basis. The survey is intended to measure the commercial and institutional sector in Canada in terms of floor area occupied, energy consumption, and other characteristics. Beginning with reference year 2009, an annual survey of establishments is conducted to support establishment-based analysis, and a building component is included on a periodic basis to support analysis of building-level data. For years which include the building component, a survey of buildings will also be conducted in parallel using the survey of establishments to produce a list of eligible buildings from which a sample is selected. The data collection is co-ordinated between buildings and establishments to reduce response burden where possible and the Weight Share Method is used to derive estimation weights for building-level analysis. This paper provides an overview of the project, and describes the methodology for several important steps in the survey within the context of indirect sampling, namely the sample design, data collection, and estimation.

KEY WORDS: Building survey, Energy consumption, Establishment survey, Indirect sampling, Weight Share Method.

RÉSUMÉ

L'Enquête sur l'utilisation commerciale et institutionnelle d'énergie (EUCIE) est réalisée chaque année par Statistique Canada au nom de Ressources Naturelles Canada et Environnement Canada. L'enquête a pour objectif de mesurer le secteur commercial et institutionnel en termes de surface de plancher occupée, l'utilisation d'énergie ainsi que d'autres caractéristiques. À partir de l'année de référence 2009, une enquête annuelle auprès d'établissements est menée pour l'analyse au niveau des établissements et une composante bâtiment est incluse périodiquement pour permettre l'analyse des données au niveau des bâtiments. Pour les années qui incluent la composante bâtiment, une enquête parallèle sera faite en utilisant l'échantillon des établissements pour produire une liste partielle des bâtiments de laquelle l'échantillon est tiré. La collecte des données des établissements et des bâtiments est coordonnée afin de réduire le fardeau de réponse, et la méthode du partage des poids est utilisée afin de dériver les poids d'échantillon des bâtiments pour l'analyse. Cette article donne un aperçu du projet et décrit la méthodologie de certaines étapes importantes de l'enquête, à savoir le plan d'échantillonnage, la collecte des données et l'estimation, dans le contexte du sondage indirect.

MOTS CLÉS : Consommation de l'énergie; échantillonnage indirect; enquête établissement; enquête bâtiment ; méthode de partage des poids.

1 RECENT STATISTICS CANADA SURVEYS ON ENERGY USE

Energy use is measured by Statistics Canada by a number of ongoing survey programs. In general, the energy consumption statistics for Canada are divided into the following three sectors: Commercial and Institutional, Residential, and Industrial. The focus of this paper is on the measurement of the Commercial and Institutional Sector, but interested readers are encouraged to consult www.nrcan.gc.ca for information on the measurement of other sectors.

1.1 Commercial and Institutional Consumption of Energy Survey (CICES 2003 – 2008)

The Commercial and Institutional Consumption of Energy Survey, an establishment-based survey, was conducted annually from 2003 to 2008 (with the exception of 2006 where no survey was conducted). This survey used the Business Register of Statistics Canada as well as some supplementary lists as the survey frame, and a total sample of 9,500 establishments was selected from a population of approximately 750,000 establishments. Data were collected via a mail-

[1] Steve Matthews, 100 Tunney's Pasture Driveway (17-D), Ottawa, Ontario, Canada, K1A 0T6, steve.matthews@statcan.gc.ca
[2] Tyler Kirkland, 100 Tunney's Pasture Driveway (16-P), Ottawa, Ontario, Canada, K1A 0T6, tyler.kirkland@statcan.gc.ca

out/mail-back questionnaire distributed after a telephone pre-contact to identify an appropriate respondent as well as confirm the establishment's eligibility for the survey (i.e., exclude home offices, out-of-scope units, etc.).

1.2 2000 Commercial and Institutional Building Energy Use Survey (CIBEUS)

The CIBEUS survey was conducted once to produce building-level energy consumption data of commercial and institutional buildings in 2000 for Natural Resources Canada. This survey used a two-stage approach; the first stage consisted of a random selection of geographic areas within which all commercial and institutional buildings were listed. The second stage involved sampling of the listed commercial and institutional buildings. Personal interviews were then conducted for each selected building. The sample included approximately 5,000 buildings and a response rate of 85% was achieved. While the survey results did meet the intended use, the survey was costly due to the listing exercise required for each selected geographic area, and the time period required to complete the project was longer than desired.

2 NEED FOR A NEW SURVEY

Historically, the key estimates provided from the CICES survey have been estimates of total floor space, total energy consumption and the overall ratio of these two quantities, referred to as energy intensity (total energy consumption / total floor space). These estimates were produced at industrial and geographic levels. In addition to these key indicators, a data-sharing agreement has been in place for Natural Resources Canada to use the micro data for ad hoc analyses such as those presented in Natural Resources Canada (2011). The need for this establishment-level data on an annual basis is ongoing. However, several improvements to the survey design of CICES were desired for the new survey program, namely an increased response rate and improved precision of estimates (for cross-sectional as well as trends over time).

There is currently a need for building-level data on energy consumption. This requirement arises from an initiative to develop baseline energy ratings for commercial and institutional buildings. The purpose of this project is to develop a predictive model to estimate the expected energy consumption of buildings based on a number of building characteristics. Individual businesses would then be able to compare their actual energy consumption to their expected energy consumption (via the model) and buildings that compared favourably would be eligible to receive an energy star rating. The building-level data will be used to develop the predictive energy consumption model.

2.1 Plan for an Integrated Survey

To address the current data needs, a number of survey design options were considered. A first option was to conduct independent establishment and building surveys. This would entail continuing the annual establishment survey, and repeating the 2000 building survey for the current reference period, but the cost of the listing exercise required under the design of the previous building survey was prohibitive. Since no other complete frame of buildings is currently available, other options to leverage the existing lists from the 2000 reference period, and develop an approach to take the changes to the population into account were investigated. In the end, it was concluded that the resulting methodology would not lead to consistent estimates over time, and the efficiency of cross-sectional estimates would deteriorate over time.

Much of the content currently required of establishments and buildings is similar or even identical. In addition, there is a large portion of the population for which the establishment corresponds exactly to a single building, so there would be some duplication of effort in carrying out parallel surveys for building- and establishment-level data. Given the similarities in the outputs of each survey, there is a desire to have some level of coherence between building and establishment-level outputs for the same reference period.

In the interest of creating a survey program that would meet the ongoing needs of both the establishment- and building-based data users, an integrated survey program was designed. Throughout this document, reference will be made to the building component and the establishment component of the integrated survey.

3. THE SURVEY OF COMMERCIAL AND INSTITUTIONAL ENERGY USE

3.1 Target Population and Sample Design

The target population for the SCIEU survey is defined as follows:

Establishment component – All establishments whose main business activity is in the commercial and institutional sector, who rent or own commercial or institutional space, and have at least one employee. Establishments located in the Canadian territories (Yukon, Northwest Territories and Nunavut) are excluded.
Building component – All buildings with at least one employee working during the main shift and at least 50% of their space being used for commercial or institutional purposes. Buildings located in the Canadian territories are excluded.

Conceptually, these two target populations represent mainly the same physical areas with some exceptions. For example, a commercial establishment located in a building that is primarily residential is in the target population for establishments, however the building is out of scope. Conversely, areas occupied by manufacturing establishments within a building that is primarily used for commercial or institutional purposes is outside of the establishment target population, while the corresponding building is in-scope for the building component. Based on results of past surveys, these cases are expected to be rare.

The survey frame was extracted from Statistics Canada's Business Register, described in Statistics Canada (2000), and supplemented with lists of hospitals as well as primary and secondary schools. The frame was created at the establishment level by excluding inactive businesses, non-employers and businesses operating outside of the commercial and institutional sectors or within the territories specified above. While this process will generate a list frame of establishments, it should be noted that no formal frame of buildings was obtained, and the sample of buildings is selected via indirect sampling methods as outlined below.

The sample design has been developed in order to produce estimates with target precision as follows:
Establishment component – estimates of proportions with 5% standard errors within each industry by region domain. For estimation purposes, the population is divided into 19 industry groups as well as 5 geographical regions (listed below).
Building component – estimates of proportions with 8% standard errors within each building-type by climate zone domain. Building type classifies each building into one of 10 categories, according to the primary type of commercial or institutional activity in the building. Climate zone divides the population of buildings into 4 distinct geographic areas (listed below) based on postal codes.

The sample of establishments is used directly to support estimates based on establishment data, and indirectly to support estimates based on building level data. This aspect is described in more detail throughout this paper, but for the sample design to be efficient, this implies that all domain variables available on the frame and identified for either establishment- or building-level estimates need to be included in the stratification of the establishment sample. In order to support the identified estimates, the frame of establishments was stratified by the following variables:

- Regions (Atlantic, Quebec, Ontario, Prairies and British Columbia)
- Climate Zone (Atlantic, St. Lawrence Corridor, Prairies and Coastal B.C.) – based on establishment postal code
- Estimated Physical Size (Small and Large) – based on proxy variables such as number of employees
- Industries (19 categories)

This cross-classification resulted in approximately 300 non-empty strata. Based on the target precision requirements, and assumed rates of out-of-scope (25%) and non-response (25%), the initial establishment sample is approximately 7,000 units. This is expected to be large enough to satisfy the establishment-level data requirements as well as the building-level data requirements (based on estimated counts of buildings of each type within each industry).

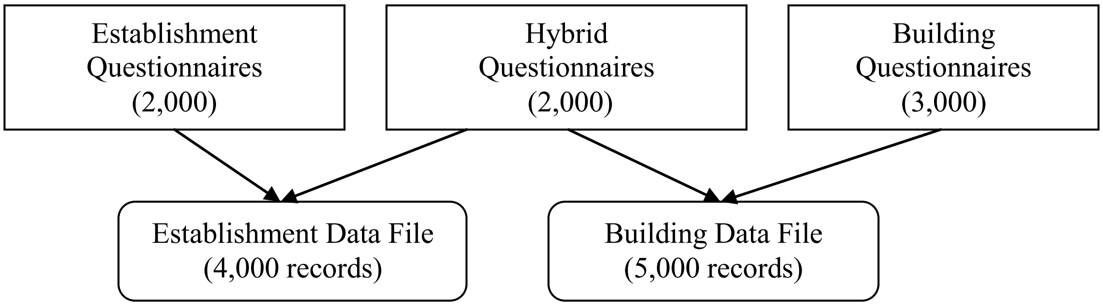
The sample of establishments is selected directly from the list frame via our stratified design. Once the list of buildings is available for each selected establishment, a sample of buildings is selected for each establishment. At least one building is selected for each establishment, and a maximum of four are selected for multi-building establishments to limit the response burden on individual businesses. Since many buildings include more than one establishment, the classical theory of two-stage sampling does not apply.

3.2 Data Collection

The data collection process consists of two components: a telephone pre-contact and personal interviews. During the telephone pre-contact, the person contacted at each selected establishment is asked questions to confirm that their establishment is eligible for the survey, as well as to provide a list of all physical buildings that they occupy (used to select the building sample), and identify a suitable respondent for the entire establishment, as well as for each listed building.

After the telephone pre-contact is complete, the sample of buildings and establishments is finalized and interviewers make arrangements to conduct computer-assisted personal interviews (CAPI) for each building and establishment that has been selected. For establishments where the building and establishment are equivalent, collection of building-level and establishment-level data are co-ordinated into one interview (referred to as a hybrid case). Operationally, the collection for establishments and buildings is managed by one collection application which identifies the relevant questions for each interview (building, establishment or both) and includes automated skip patterns to guide the interviewer to ask only the proper questions of each respondent. The CAPI application also includes a number of edits to validate the reported data.

After accounting for some non-response and out-of-scope units among the initial sample of 7,000 establishments, it is expected to have reported data for approximately 4,000 establishments (establishment and hybrid questionnaires combined). The combination of building level questionnaires and hybrid questionnaires (which also report building level data) is expected to result in reported data for approximately 5,000 buildings. The diagram below shows the assumed breakdown in terms of the number of completed questionnaires for each type, as well as the number of records available for analysis on the final data files. These anticipated counts are approximately equal to what was achieved for the 2009 survey.


Establishment Questionnaires (2,000)
Hybrid Questionnaires (2,000)
Building Questionnaires (3,000)
Establishment Data File (4,000 records)
Building Data File (5,000 records)

3.3 Treatment of Item Non-Response

Item non-response is treated by imputation which is carried out using the Banff system, Statistics Canada's generalized imputation software which is described in Banff Support Team (2008) . Banff provides procedures to impute for non-response using a nearest-neighbour donor approach that can be customized for each question (or group of questions). The approach is customizable in the sense that, using parameters in the Banff processor, specific records can be excluded, and the imputation groups and matching fields that are used to match donors to recipients can be specified. The questionnaire is broken into 51 distinct blocks, and the imputation is carried out independently within each block. Within each block, if a record requires imputation for one or more cells, all cells that require imputation are copied from the selected donor record.

3.4 Estimation

In producing establishment-based estimates, a multi-phase weight is used. The first component of the weight is based on the selection probability in the sample design, the second component accounts for unit non-response from pre-contact among the sampled establishments, and the third component represents a weight adjustment for non-response during the personal interviews. This weight allows the set of respondents to be used for valid inference on the population of establishments, assuming that the non-response mechanism within each stratum is ignorable. Statistics Canada's Generalized Estimation System (GES) is used to produce the survey estimates. Information on GES can be found in GES Support Team. (2005).

For building-based analysis, the Weight Share Method (WSM) is used to derive weights. The establishment weights which have already been adjusted for non-response are used, as well as information on the relationships between buildings and establishments as outlined below. A further weight adjustment is applied to these initial building weights to take into account the unit non-response among buildings. This can be seen as a two-stage design where the first stage incorporates the sample selection and non-response at the establishment level and the second stage comprises the non-response and sub-sampling at the building level. However, standard two-stage methods such as a cluster design do not apply since the relationships between establishments and buildings can be quite complex (i.e., establishments do not comprise mutually exclusive groups of buildings). For example, some buildings may be occupied by more than one establishment in the population, and this needs to be taken into account as it affects the probability of selection of the buildings. A comprehensive summary of the theory and application of the WMS method, including the generalized form is provided in Lavallée (2002).

The formula that is used to derive estimation weights for the buildings is:

$$w_j^{BLD} = \frac{\sum_{i \in s_E^{PC}} \left(w_{12i} \bullet w_{4i}^{SUBSAMP} \right) \bullet I(i \supset j)}{\sum_{i \in U_E} I(i \supset j)} \bullet w_{3j}^{NR(INT)}$$

The index of the building (the indirectly sampled unit) is denoted by *j*, and i is used to denote the index of the establishment (the directly sampled unit). Here, s_E^{PC} is the sample of establishments for which pre-contact was completed, w_{12i} is the weight of establishment i discussed further below and $w_{4i}^{SUBSAMP}$ is the weight associated with the sampling of buildings from establishment *i*. As well, $I(i \supset j)$ is an indicator variable equal to 1 if establishment i is linked with building j and equal to 0 otherwise and U_E represents the population of establishments. The term $w_{3j}^{NR(INT)}$ is a weight adjustment described in more detail below.

In applying the WSM for SCIEU, three design aspects specific to the survey were taken into account:

- Since non-response at the establishment level is encountered during pre-contact, the fact that the pre-contact is not able to be completed and the building roster is not collected for all selected establishments needs to be accounted for. It is assumed that the response mechanism at pre-contact is uniform within stratum, and the resulting sample of completed units is treated as a stratified sample from the frame when conditioning on the number of completed units. Thus, the numerator of the building weight is based on this reduced sample, denoted s_E^{PC}, consisting of the set of all sampled establishments for which pre-contact was completed. The corresponding weight that is applied is $w_{12i} = w_{1i}^{Design} \bullet w_{2i}^{NR(PC)}$, which is the establishment design weight, adjusted for non-response at the establishment level during pre-contact.
- For establishments with 4 or more buildings, a second stage of sampling is applied for the selection of buildings. This results in a corresponding weight $w_{4i}^{SUBSAMP} = 1 / f_i$ that is included in the summation in the numerator of the building weight, where f_i is the sampling fraction of buildings for establishment *i*.
- Finally, non-response is encountered at the building level during the personal interview stage. Again, it is assumed that the non-response is generated uniformly within each stratum, and the building weights are ultimately adjusted by a factor of $w_{3j}^{NR(INT)}$ so that inferences can be drawn for the entire population of buildings.

It is clear from the formula that information on the relationships between buildings and establishments, $I(i \supset j)$, is required for each sampled building in order to calculate the estimation weights. The numerator is the sum of the weights for all sampled establishments that occupy the building. The linkage indicators are available from the building rosters collected at pre-contact for all responding establishments. The denominator of the weight is the count of establishments in the survey population that occupy the building. For this quantity, any sampled establishment that includes the building in their roster, and any non-sampled establishments that occupy the building according to an administrative source (Statistics Canada's Business Register) are counted.

Considering this weight further, if the non-response adjustments are disregarded, it can be seen that two scenarios are possible:

- For buildings that are occupied by a single establishment (that is, $\sum_{i \in U_E} I(i \supset j) = 1$) the building weight will be analogous to a two-stage weight with the appropriate sampling fraction at the second stage.
- For buildings that are occupied by multiple establishments, the WSM will produce a weight that takes into account the multiple establishments that, if selected, would have lead to the inclusion of the building. This weight will not be equivalent to the weight that would be calculated by considering the exact inclusion probabilities.

This methodology will lead to non-negative weights, but in some cases the weights can be less than one. This occurred for approximately 8% of the buildings for the SCIEU survey. To avoid issues of interpretation for weights less than one, an adjustment was made to the weights so that each weight was at least one, and the weights of other buildings within the stratum were adjusted to yield the population counts.

Point estimates and variance estimates were produced using Statistics Canada's Generalized Estimation System (GES) which allows us to specify domains of estimation and produce estimates of proportions, totals and ratios. To implement the weights calculated by the WSM in the stratified framework, the resulting estimator is expressed as one based on a cluster design where establishments are treated as clusters of buildings. Under this approach, $\hat{Y} = \sum_{j\in s'_B} w_j^{BLD} y_j = \sum_{i\in s'_E} w'_i \left[\sum_{j\in s'_B} w'_{j(i)} y_j \right]$, where, s'_B represents the set of all responding buildings, $w'_i = N_h / n_h^{s'_E}$ is the adjusted weight of the establishment where s'_E is the set of establishments within the sample that are linked to at least one responding building, and $n_h^{s'_E}$ is the number of such units within stratum *h*. As well, $w'_{j(i)} = w_j^{BLD} \bullet \left(w_i \Big/ \sum_{i\in s'_E} w_i I(i \supset j) \right)$ if establishment *i* occupies building *j* and $w'_{j(i)} = 0$ otherwise. This expression allows for a given building to be represented within more than one cluster if multiple establishments within the building are selected. The corresponding variance estimates properly take the sample design into account, with the exception of the sub-sampling of buildings for establishments with 4 or more buildings. Since this is not common, it is expected that the resulting underestimation of variance will be minor. For further details on applying the WSM estimator in the context of a stratified simple random sample without replacement, please refer to Beaumont (2007).

4. FUTURE PLANS

The establishment component of SCIEU is expected to be conducted annually, while the building component will occur on a less frequent basis, likely every 4-6 years. For the next iteration of the establishment survey, a number of longitudinal aspects will be incorporated into the design, including increased sample overlap, collection strategies to encourage repeat respondents, and estimation and imputation methods that make use of historical data. These enhancements are intended to produce more efficient estimates of year-over-year trends. As well, variance estimation that includes a measure of variance due to imputation for item non-response will be developed. For the next iteration that includes a building component, the variance estimation procedures required to include the component of variance associated with the sub-sampling of buildings, as well as non-response and imputation will be developed.

REFERENCES

Banff Support Team. (2008). "Functional Description of the Banff System for Edit and Imputation". *Statistics Canada Technical Report.*

Beaumont, J.F. (2007). "A note on Weighting and Estimation in the Film Production Survey". *Statistics Canada Internal Document.*

GES Support Team. (2005). "GES v4.3 Overview". *Statistics Canada Technical Report.*

Lavallée, Pierre (2002). *Le Sondage Indirect, ou la méthode généralisée du partage des poids*. Éditions de l'université de Bruxelles, Brussels.

Natural Resources Canada (2011). "2008 Commercial & Institutional Consumption of Energy Survey – Summary Report", M141-17/1-2008.

Statistics Canada (2009). "A Brief Guide to the BR". *Statistics Canada Internal Document.*