ENVIRONMENTAL STATISTICS AT STATISTICS CANADA

Claude Girard, Martin Hamel and John Marshall[1]

ABSTRACT

Environmental surveys are very recent additions to Statistics Canada's economic program. Carrying such surveys, as it was discovered, raises some very interesting and unique challenges. The paper presents some of these issues and discusses some of the work done so far to address them.

KEY WORDS: Environmental Surveys, Two-Phase Strategy.

RÉSUMÉ

Les enquêtes environnementales sont de toutes récentes additions au programme d'enquêtes économiques menées par Statistique Canada. Comme nous en avons fait l'expérience, mettre en œuvre des enquêtes ayant pour sujet d'intérêt l'environnement pose des défis intéressants et, dans certains cas mêmes, uniques. Nous présentons dans ce document quelques uns de ces enjeux, ainsi que les solutions que nous avons considérées jusqu'à présent.

MOTS CLÉS : Enquêtes environnementales; plan d'échantillonnage à deux-phases

1. INTRODUCTION

This paper is about the environmental surveys program at Statistics Canada. It provides an overview of the program, along with some of the methodological and analytical issues that were encountered conducting environmental surveys, which are recent additions to the business surveys program.

The paper is structured as follows. First, the background behind the development of Statistics Canada's environmental surveys is presented - what they are, why the information is required, and what are some of the challenges collecting it. Secondly, the paper describes the methodological context set out by economic surveys under which the newer environmental surveys were introduced. The lack of a perfect fit between these two types of surveys has given rise to methodological issues that are discussed in the paper. Finally, concluding words on the lessons learnt throughout are given.

2. BACKGROUND

In Canada, very limited environmental information was collected before the 1980s. In fact, the first dedicated environmental survey at Statistics Canada was not launched until the late 1980s when the Federal Government introduced the "Green Plan for a Healthy Environment". This was the key turning point for the future development of environmental surveys. Since then, a number of other programs have been put in place at the federal level starting in the 1990's, which have allowed for the development of a "suite" of environmental surveys.

What began as a program responsible for conducting two surveys, one profiling the waste management industry (business and government) and the other environmental expenditure in the business sector, has grown into one including numerous business surveys on a variety of topics, as well as two household-based surveys. This growth has largely occurred because of increasing interest on the part of other Federal Departments. The need for robust environmental data has allowed the program to continue to grow. Environmental statistics are considered to be a priority by many, both inside and outside of the Federal Government, and this is illustrated best by the wide range of clients who use the survey information, including:

1 Statistics Canada, Ottawa (Ontario), K1A 0T6; claude.girard@statcan.gc.ca, martin.hamel@statcan.gc.ca, john.marshall@statcan.gc.ca

- Federal government departments and agencies – most notably, Environment Canada, Natural Resources Canada, Health Canada, Industry Canada, Agriculture and Agri-Food Canada, Canada Mortgage and Housing Corporation, Health Canada and the National Research Council;
- Provincial and Territorial governments;
- Municipal governments; Universities and non-governmental organizations (NGOs);
- Traditional and social media.

The data are disseminated using a number of different Statistics Canada publications, or through reporting mechanisms and requirements of international and national organizations. Internally, all surveys are required to produce an analytical survey report that covers all or selected elements of the data that were collected. The methodology and concepts used in the survey are also clearly described.

Data are also published using one of the following vehicles:

- EnviroStats – quarterly publication for a general audience;
- Human Activity and the Environment – annual statistical compendium with an analytical article focussing on a selected environmental topic.

Externally, data are used for the development of national, provincial and local environmental indicators, market research and in academia, among other applications. There is also a significant international reporting aspect, as many of the statistics are included in international environmental data compilations such as the Organisation for Economic Cooperation and Development (OECD) Joint Questionnaire.

There are four established surveys:

- Survey of Environmental Goods and Services (replacing the original supply-side Environment Industry Survey);
- Survey of Environmental Protection Expenditures;
- Waste Management Industry Survey: Business Sector;
- Waste Management Industry Survey: Government Sector.

These surveys formed the initial "core" of the environment statistics program at Statistics Canada. In addition to measuring different environmentally-related elements of the economy, they filled a data gap which existed in this subject area in the System of National Accounts.

A number of newer surveys have recently been added to the program:

- Industrial Water Surveys;
- Survey of Industrial Processes (pilot);
- Households and the Environment Survey;
- Household Energy Survey;
- Survey of Drinking Water Plants;
- Agricultural Water Survey;
- Farm Environmental Management Survey.

The motivation and sources of funding for these surveys are more diverse than the original four, reflecting the increased awareness and importance attached to collecting reliable data in profiling and studying the effect of human activity on the environment.

3. METHODOLOGICAL PERSPECTIVE

The methodology behind Statistics Canada's business environment surveys is handled by a team of methodologists working on business surveys (consisting almost exclusively of economic surveys), in concert with a team from the subject matter area.

In order to maintain a high level of quality for its (growing) economic program, while not spending more resources and money than it has to, the Agency strongly encourages the use of proven and well-tested methodologies in all surveying activities.

This is the context in which the emerging environmental surveys operate under: All surveys - existing and recent additions to the program alike - must make the best possible use of the existing resources. One issue this brings, and this is the general theme behind this paper, is that existing methodologies and systems may not be always well-fitted to environmental surveys, which often present their own specific challenges.

Each of the following sections describes an issue that can arise with environmental surveys, along with the steps taken so far toward a satisfactory solution; there is one section for each of the following key methodological steps of a survey:

- Frame creation;
- Sampling;
- Edit and Imputation; and,
- Release of statistical information.

3.1 Issues related to the frame

At Statistics Canada, business surveys get their frame from the Business Register (BR). To give a comprehensive description of the BR would go beyond the purpose of this paper; the interested reader is referred to Statistics Canada (2010). Essentially, the BR is a vast centralized database seeking to encompass all of the business economic activity that is taking place in the country. It is a database containing the most comprehensive list of all business enterprises in Canada. It gets its information from basically three sources:

- The Canada Revenue Agency (CRA);
- Survey feedback;
- In-house quality surveys.

There are at least three ways to represent the structure of a given corporation: 1) How the corporation sees itself and its sub-entities; 2) How the CRA sees it; and 3) How the BR represents it, using what are called “statistical entities”. Based on the information obtained by the BR, the corporation is partitioned by grouping its sub-entities into one of four nested statistical entities; they are, by descending order of size: enterprise, company, establishment and location. So, the enterprise is the broadest level within a corporation whereas the location is its most specific. There may be several locations attached to a single enterprise; these are known as complex corporations.

Typically, economic surveys do their sampling at the establishment level. In contrast, many environmental surveys elect to do their sampling at the location level: the level at which the sought-after information is most-likely to be found.

Given its size, keeping the entire BR continually up-to-date is a huge and complex undertaking. Consequently, priorities have to be set and these reflect the needs of the most important and common BR users: the economic surveys. For instance, survey feedback is an automated post-collection process that informs the BR of any change in status a survey has encountered about a given unit during collection (e.g., a change in main business activity). But because economic surveys do not sample entities which are assumed to be small in terms of revenue as frequently as larger units (as a means to alleviate response burden, for instance), information about small entities do not get updated as regularly as larger ones. However, environmental surveys, unlike their economic counterparts, are not necessarily ready to exclude small entities from their surveying activities: small units *may* prove to hold a significant portion of the activity these surveys are primarily interested in. Also, establishment-level information is usually more reliable than location-based information, mainly because the latter is seldom used by economic surveys.

3.1.1 The example of the Survey of Industrial Processes

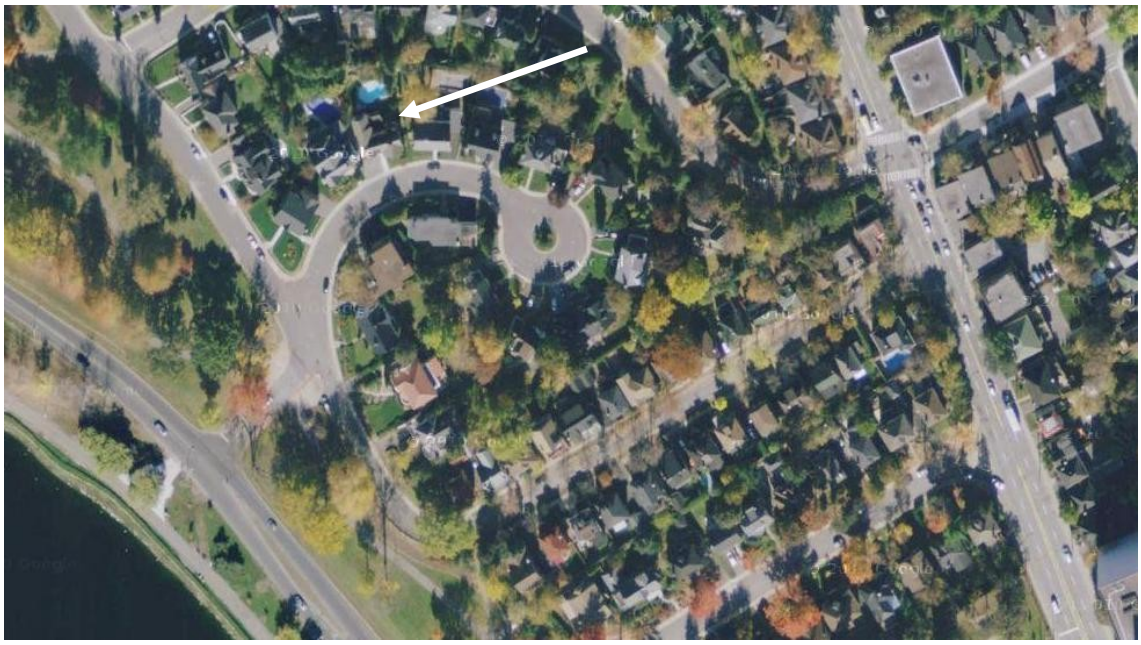
The Survey of Industrial Processes (SIP) is a novel initiative at Statistics Canada, introduced as a possible solution around an issue many environmental surveys are facing: the concepts behind the data to be collected can be very technical in nature.

Through its pilot component that focuses on Retail Gas Stations (RGS), the SIP investigated the possibility of estimating benzene emissions that retail gas stations produce in their daily activities, by collecting from gas stations operators what is at the root non-technical information. More specifically, information is sought about the processes used in running a retail gas station which are known (or suspected) to result in benzene emissions. The information collected is then turned into emissions data using engineering models gathered and/or developed by the subject matter area. This avoids having to ask operators to provide us with their own estimate of on-site emissions, something they may not be able to do given that expert technicians with expensive equipment would have to be called in to assess the situation.

As an example of process information, the SIP questionnaire asks about the number, type and age of the tanks found on site, and also about the number of fuel deliveries that take place in a given week. These are only some of the factors behind on-site emissions. The general idea is the following: the greater the number of times the fuel containment seal is broken at a given site, the larger the quantity of benzene emissions. Indeed, in principle, if the gasoline containment system was water-tight from one end to another (i.e., from its delivery and storage to its distribution to a car's tank), then no emissions would ever occur. But, in practice, despite the equipment used, emissions *do* occur at each step. For instance, the seal gets broken on the storage tank when a fuel delivery takes place (i.e., when the tank cover is lifted), despite the fact that a vapour recovery device is being used. Another source of emissions is when a customer actually fills up his car. In both of these cases, the familiar smell of gasoline one experiences when visiting an outlet is a sign of benzene emissions.

The SIP focuses on *all* retail gas stations, big and small revenue-wise, and at the finest level found on the BR: the statistical location entity. The initial frame extraction from the BR identified about 20,000 *records* for the retail gas station industry (according to North American Industrial Classification System (NAICS) codes). Subsequent research on the Web revealed that the Calgary-based consulting firm MJ Ervin has been conducting a census of retail gas station *outlets* every two years. Their most recent census claims that there were in 2008 about 13,000 *outlets* in Canada, which is significantly lower than the frame count of the statistical entities that are the BR *locations*.

A coverage study of the SIP frame was undertaken to better understand the discrepancy between the two. It had to take place with minimal burden on respondents. It was accomplished by performing "virtual visits" using Web technologies for a sample of BR records. More specifically, a number of frame records were drawn and GoogleMaps SatelliteView® was used to determine the nature of the physical entity tied to the location. (At the time, only the SatelliteView® mode was available in Canada; the much more detailed StreetView® came in effect only later.) The idea was that a retail gas station has well-recognizable features even when seen from above. To illustrate, the following images are satellite views corresponding to two records on the SIP frame: the first clearly is a retail gas station, while the second (indicated by the arrow), is not. (A pool is clearly present in the backyard of this residential property located in a cul-de-sac).



Such findings make it clearer what precisely can be found on the SIP frame. For instance, records pointing to a residential area could presumably be associated with a real outlet through, say, the accountant's address or the owner's address as shown on the corresponding CRA record.

The Web, especially since the StreetView® has come in effect, offers new "burden-free" ways to validate the information contained in a business survey frame, since fewer (or even no) calls or contacts are needed. This is a worthy option to consider for any survey interested in well-recognizable physical entities like retail gas stations, mills, etc. But great care must be exerted, however, with respect to the information contained in these views since it is gradually becoming outdated: How often, if at all, will a given area be re-visited by Google (or similar Web entities) in the future?

3.2 Issues related to sampling

Efficient sampling of business surveys relies on a strong concordance between the units one *needs* to have on file and those one *manages* to get on file. On the BR, a business is represented through its main economic activity, which is what drives its NAICS designation. The issue with environmental surveys is that the activity of primary interest may not necessarily be the main business activity as identified on the BR. In other words, the "front window" NAICS may not show a given environmental activity, although it may indeed be taking place there. As a result, a NAICS-based extraction of the frame from the BR has to cast its net wide. Typically, it will admit any NAICS which are known (or simply suspected) to have units involved in the sub-activities in which the environmental survey is interested. This may result in a frame presenting a much higher out-of-scope rate than other business surveys. This issue, which may arise also for economic surveys interested in data about commodities, is amplified to some extent in environmental surveys which tend to focus on highly specialized sub-activities.

3.2.1 The example of the Survey of Environmental Goods and Services

The Survey of Environmental Goods and Services (SEGS) focuses on units that produce certain specific environmental goods and/or services. The frame one can get from the BR at the most detailed level industry-wise is made of units pertaining to a pre-determined list of NAICS-6. While these selected NAICS-6s are known to harbour the sought-after units, they also include lots of units in which SEGS has no interest in.

Consider NAICS-6 = 322220 comprising units that manufacture yard waste bags, which are units of interest to SEGS. This one NAICS-6 also contains units engaged in any one of the following (non-exhaustive) list of activities:

- Adhesive tape (except medical);
- Cardboard (made from purchased paperboard);
- Gift wrap, laminated, made from purchased paper;

- Wallpaper;
- Waxpaper.

None of these activities, however, is of interest to SEGS. The ratio of units of interest to SEGS to the total number of units in a given NAICS is called a prevalence rate. Because the prevalence rate per NAICS-6 was not known prior to the first edition of SEGS, it was the subject matter area's responsibility to come up with educated guesses, which were needed to assess the size of the sample needed to fulfill the survey's needs. As it turned out, depending on the NAICS-6, the prevalence rate could be as low as 5% for certain NAICS-6 groups, with one whose rate was 100%.

The traditional one-phase sampling strategy employed by economic surveys is not appropriate for SEGS's needs, as it would entail a significant loss of efficiency due to the high anticipated out-of-scope rate. The approach adopted for SEGS is a two-phase sampling design, which is the textbook solution for such a situation. While this may be true, implementing a two-phase strategy in a national statistical agency like Statistics Canada does raise some significant issues of its own.

The first-phase of SEGS involved contacting over the phone a large sample of units drawn from the BR within the list of pre-determined NAICS-6 groups (as identified by the subject matter area). The short set of questions, which should have been technical and exhaustive to suit SEGS' purpose adequately, were kept simple to make the interview run more smoothly and to reduce response burden. The respondent was basically asked whether the unit is involved in the manufacture or import of broad categories of environmental goods. Because the questions were simple, they did not allow to fully zero-in on units of interest: some units that perceived some of their activities as fitting the bill were not in fact qualifying, and thus ended up being excluded from the second-phase sample.

Even though the questions were deemed simple, the person identified on the BR as a contact person often was not able to provide us with a "simple" answer. Indeed, such a person may be a clerk or an accountant - someone authorized to answer the usual economic-type queries, but not environmental questions. Consequently, the initial phone call often morphed into fax exchanges, a format many respondents found more suitable, possibly because it allowed the questions to be transferred to their on-site expert. These additional contacts have led to more time and resources needed to carry out the first phase of SEGS collection.

Even though conceptually a first-phase questionnaire should be simple to answer, many respondents may perceive the two-phase strategy as two separate and "regular surveys". This is such a strong perception that some survey areas where respondents are already burdened will not consider a two-phase approach even if, on paper, it addresses their sampling needs.

Generalised collection systems do not easily accommodate a two-phase sampling strategy, which calls for a "break" during collection, at the point when second-phase sampling is to begin. It takes a lot of flexibility in the collection systems to accommodate all the situations that can arise. Fortunately, SEGS did not require much in the way of modifications between the two phases, and the "break" really was just that: a pause. The collection process was "resumed" after phase one without much change to the files involved. With hindsight, the second phase presented an opportunity to adjust the sample, should the observed prevalence rate of units of interest within SEGS' NAICS-6s differ significantly from the educated guess initially made to create the sample (i.e., finding *too many* units of interest would have compromised the allocated budget and so a second phase of sampling would have been required to keep the number of questionnaires to be sent to an acceptable level budget-wise.)

While the two-phase was and remains one solid solution to the issues SEGS raised, it is worth rising, as it was done here, that practical issues need to be overcome for this textbook solution to be fully beneficial.

3.3 Issues related with Edit and Imputation

Edit and Imputation, commonly referred to as E&I, is a key survey step where quality of the data collected is closely monitored and improved upon, if possible. This is a delicate step which requires both knowledge of the subject and expertise of the methods used. While expertise of the methods is transferable from economic to environmental surveys, knowledge of the subject obviously is not. This is an issue that all new surveys (or areas) face.

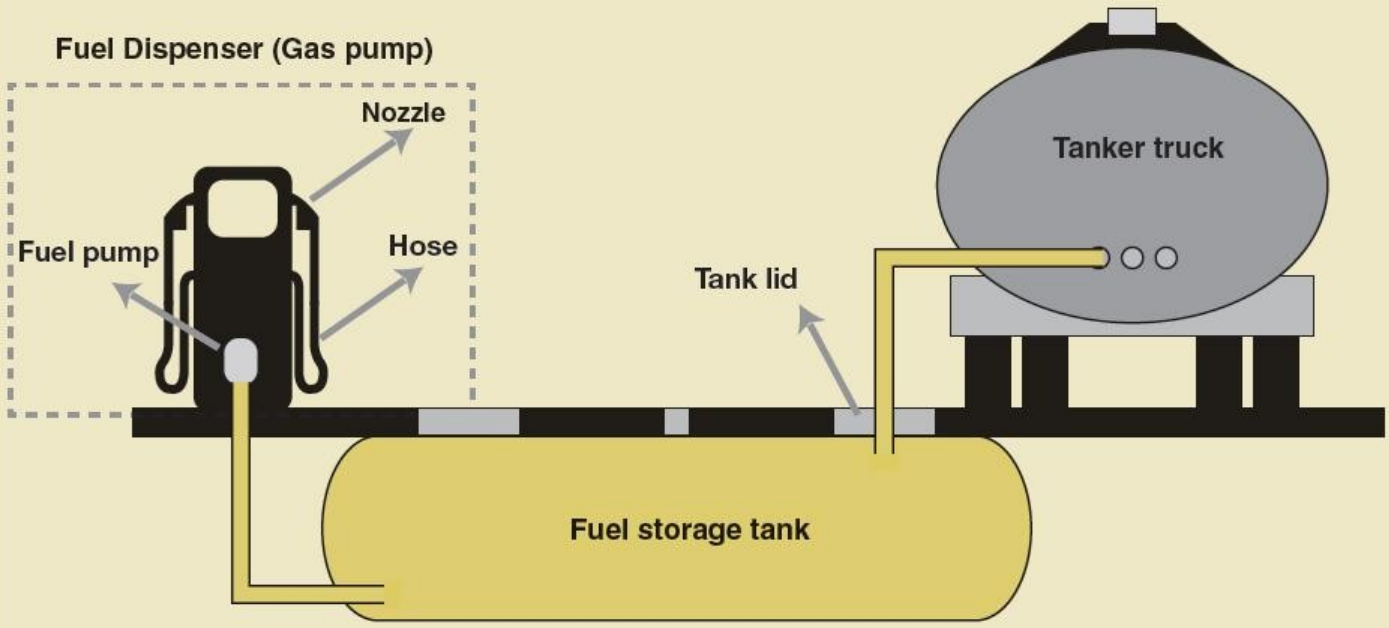
In the context of the SIP, an innovative approach to E&I was used that allowed an acquisition of "on the fly" knowledge which was lacking at the beginning of the imputation process. Specifically, the fact that retail gas stations are easily

recognized physical entities, whose main processes and attributes can be witnessed through an on-site visit, was used to corroborate information submitted by respondents. The use of Google StreetView® enabled *virtual* on-site visits to settle questions that arose about the data just collected.

3.3.1 The example of the SIP

As an example, the SIP was interested in the number of dispensers there are on site. SIP's questionnaire contained a detailed definition of what was meant by the term "dispenser".

Reporting guide



Example of industrial processes in a typical retail gas station

Section 5 – Fuel Dispensers

- **Dispenser:** An apparatus to pump gasoline, diesel or other vehicle fuels. A dispenser may have more than one nozzle serving multiple clients. A dispenser may have several "pump numbers" to identify clients for payments.
- **Hand-held automatic shut-off nozzle:** A nozzle designed to help prevent spills during refuelling by customers. Such nozzles have a pressure sensitive mechanism that forces the release of the dispensing lever once the vehicle tank is full.
- **Hands-free nozzle with a locking lever:** A nozzle that allows the attendant, mainly in a full-service gas station, to dispense fuel without holding the nozzle during vehicle refuelling. These nozzles are equipped with a manual mechanism that locks the dispensing lever in position once pulled during refuelling of vehicles. These nozzles keep hands free while fuel is being pumped into vehicles and shuts off the flow of fuel once the vehicle tank is full.
- **Nozzle with a splash guard:** A nozzle with a round rubber guard placed on it to guard against a back splash of fuel during a fill up.
- **Nozzle with vapour recovery:** A nozzle equipped with a vacuum system that sucks escaping fuel vapours during vehicle refuelling.

Despite providing these instructions, several responses suggested that the notion of a dispenser was not always understood- see the example below.

12. How many vehicle fuel dispensers are present at this retail gas station, including both gasoline and diesel? *(Please see definition on page 3).*

C1021 1 2 vehicle fuel dispensers

For a record providing such an answer, a virtual visit was often possible. The StreetView® image below is an example of a retail gas station operating in the United States similar to what would correspond with the answer given above. It clearly has 6 dispensers, as shown by the ovals, in contrast to the 12 dispensers stated in the questionnaire. Most likely the respondent counted each side of the dispenser as a separate dispenser. (It is common in a retail gas station to have the sides of dispensers numbered 1 to *n* to help customers indicate to the clerk, from the inside, at which pump their car is parked.)

Using the Web in this manner provided a response burden-free way to validate SIP data, reducing the amount of follow-up calls to respondents.

3.4 Issues with the release of statistical information

Cell suppression is the method of choice of business surveys at Statistics Canada to prevent aggregated estimates in quantitative tables from revealing information about individual respondents. The highly sensitive nature of economic data, coupled with its potentially skewed distribution (implying that only a few units are responsible for the bulk of the value of an activity), gives rise to situations where data must be suppressed in order to respect the confidentiality clause of the Statistics Act.

The general idea behind cell suppression is illustrated by the following two tables. The tables below depict the same estimates with respect to two domains at different points in the suppression process. In the first table, certain cells are identified as sensitive (based on an arbitrary rule for the purpose of the example) and their values have been replaced by the empty-set symbol. This is the primary suppression. But since suppressing the value is not enough to prevent a user from deducing it using the remaining values and applying some basic arithmetic, additional non-sensitive cells are suppressed for additional protection. This process is known as residual suppression. The result of the primary and residual suppression is depicted by the cross signs in table 2. The lesson is clear: while primary suppression may not necessarily undermine the analytical potential of a table, residual suppression has the potential to seriously limit the amount of data disclosure.

	DOMAIN 2								
DOMAIN 1	17	23	35	**Ø**	7	14	4	1	…
	5	**Ø**	23	32	13	56	34	12	…
	56	6	45	35	12	3	**Ø**	56	…
	2	7	56	74	**Ø**	24	13	17	…
	5	**Ø**	**Ø**	53	67	45	2	5	…
	8	24	32	21	**Ø**	62	23	7	…
	12	13	16	15	4	1	**Ø**	21	…
	7	31	12	**Ø**	13	19	22	**Ø**	…
	…	…	…	…	…	…	…	…	…

Table 1: Main cell suppression

	DOMAIN 2								
DOMAIN 1	**×**	23	35	**×**	**×**	14	**×**	**×**	…
	5	**×**	**×**	**×**	13	56	34	12	…
	56	**×**	45	35	12	**×**	**×**	**×**	…
	×	7	56	74	**×**	24	**×**	17	…
	×	**×**	**×**	**×**	67	45	2	5	…
	×	**×**	32	21	**×**	62	23	**×**	…
	12	13	**×**	**×**	**×**	**×**	**×**	21	…
	7	31	**×**	**×**	**×**	19	22	**×**	…
	…	…	…	…	…	…	…	…	…

Table 2: Residual cell suppression

The nature of some environmental data, however, differs from economic data with respect to its sensitivity. For example, the thermal component of the Industrial Water Survey (IWS) collects data on the use of water from respondents which are either public or quasi-public organizations. Consequently, the information collected is not particularly sensitive in nature and most can actually be retraced through public-access requests (if not from the Web directly). Nonetheless, under the Statistics Act, the confidentiality of this information obtained through survey channels must be protected just the same.

In most surveys, cell suppression is minor since usually there are enough contributors to a given value to ensure sufficient noise, protecting any one individual contribution. However, in the thermal component of the IWS, the small number of units available for sample coupled with the disproportionate contribution to the estimates of some units, result in a large number of cells being suppressed.

In the 2007 edition of the thermal component for example, over 50% of the cells initially considered for release were suppressed, thereby significantly compromising the analytical potential of the data collected.

4. ANALYTICAL PERSPECTIVES

Identifying the target population is difficult, given that no reliable relationship exists between environmental data and standard economic measurements. Analysis has shown inconsistent relationships between environmental protection expenditures (capital and operating) with overall business expenditures, employment size, or revenues. This fact complicates the process of frame definition, which ultimately has an impact on sample selection and potentially on final estimates and data quality.

Even after several survey iterations, it is sometimes difficult for respondents to provide estimates for some variables. For example, it is sometimes difficult to report on such variables as environmental protection expenditures when the respondent is unsure as to whether a specific expenditure should, or should not, be considered ”environmental”. Another example may be the necessity of a manufacturer estimating their industrial water use and discharge when no metering exists on site, so an exact measurement is unavailable.

In spite of these challenges, much progress has been made. The business survey “infrastructure” has been fairly effective in incorporating needs of environmental surveys. Methodologists, corporate collection divisions, collection system designers are beginning to better understand the unique needs of environmental surveys. From the respondents’ perspective, businesses are becoming more familiar with environmental concepts, such as pollution prevention and environmental management.

From an analytical and/or subject matter perspective, the loss of publishable data due to confidentiality constraints is *very* important. This represents a loss of value-added for each project. It is important to see surveys as investments in information and each unpublishable cell decreases the return on investment. However, Statistics Canada is currently working on solutions that respect the confidentiality provisions of the Statistics Act while allowing more data to be published.

Finally, it is worth noticing that businesses are often keen to report their environmental protection efforts. For them, it is good public relations for an industry to be able to say that they report their environmental performances (for some measures) to Statistics Canada. This eagerness to report is reflected in the high response rates seen by these surveys - consistently between 75% and 85%. At the same time, this raises the issue of response bias, favouring “positive” outcomes, a bias that is difficult to assess and, thereby, to correct.

5. CONCLUSION

From a statistical point of view, progress in the collection and estimation of environmental data is being made on a number of fronts. But, from an analytical and subject matter perspective, some outstanding issues remain.

The issues presented here make it clear that environmental surveys cannot be conducted exactly as if they were economic surveys. From frame creation to the release of statistical information, environmental surveys exhibit some unique features. While efficiency considerations require that proven methodologies and systems developed based on years of conducting economic surveys be used as much as possible, the specificities of environmental surveys cannot be ignored. Consequently, a non-negligible amount of time and resources has to be put into adapting the current methods and systems, and into developing new ones as required.

REFERENCES

Statistics Canada (2010). *A brief Guide to the BR*. Business Register Division, Statistics Canada, Ottawa, 9 pages, July 2010.