# The Migration of the Canadian Census of Agriculture to an Integrated Business Program Without Contact with Respondents

Mathieu Thomassin[1]
Statistics Canada, Ottawa, Canada; mathieu.thomassin@canada.ca

**Abstract**

Since 1956, the Canadian Census of Agriculture (CEAG) has used a collection model based on the complete enumeration of farms and on data reported by respondents. However, this environment is rapidly changing.

Farms have become increasingly integrated and complex businesses. These businesses are best handled using Statistics Canada's business survey processing infrastructure, rather than the traditionally used social survey processes. Additionally, data requirements are becoming progressively more complex and include linkages beyond the primary production sector. Information from CEAG must be integrated with data from other sectors of the economy (e.g., the environment, food manufacturing, energy, transport, international trade and prices) to measure program efficiencies and to identify broad issues affecting one or more sectors.

The ability to complete the CEAG online has reduced the burden imposed on farm operators. Now, the increasing availability of administrative data and satellite imagery gives Statistics Canada the opportunity to eliminate all or almost all contact with agricultural producers by 2026. This would significantly reduce collection costs and preserve the level of detail and quality of information required by stakeholders. This modern approach is being implemented as a response to changes in agricultural businesses and stakeholders.

The new model will be implemented by combining remote-sensing and geospatial information, data from approximately 300 available administrative sources, data from other harmonized business surveys, and data from the introduction of models. Other non-traditional alternative sources of information, such as web scraping or precision agriculture, will also be considered.

This new model will be deployed progressively with the 2021 CEAG. A proof of concept will be produced using the new model by predicting all the census variables (nearly 200 variables) for the whole population (close to 190,000 units). In addition, up to 10 questions will be replaced by alternative data in the 2021 CEAG, using an agile collection instrument that allows data to be "smartly replaced" when the alternative source is of sufficient quality. The objective is to reduce the response burden by 100,000 hours for the 2026 CEAG.

This paper describes the long-term strategy adopted by Statistics Canada's Agriculture Division to implement its vision—CEAG-0—with the ultimate goal of eliminating all or almost all contact with agricultural producers. It also discusses how administrative data will be used for the 2021 CEAG and the challenges the team is facing.

Keywords: Administrative data; alternative data; smart replacement; response burden.

1. Disclaimer: This paper is released to inform interested parties of research related to the Canadian Census of Agriculture and to encourage discussion. The views and opinions expressed by the author should not be construed as those held by Statistics Canada.

# 1 Introduction

Since 1956, the Canadian Census of Agriculture (CEAG) has used a collection model based on the complete enumeration of farms and on data reported by respondents. The ability to complete the census online has reduced the burden imposed on farm operators. Now, the increasing availability of administrative data and satellite imagery gives Statistics Canada the opportunity to eliminate all or almost all contact with agricultural producers by 2026 (i.e., the CEAG-0 project). This would significantly reduce collection costs and preserve the level of detail and quality of information required by stakeholders. This vision is in line with Statistics Canada's modernization agenda and the innovative approaches routinely applied by the agency.

The second section of this paper provides a general description of the CEAG and introduces how administrative data are being used in the Canadian agricultural program. It also identifies how a combination of remote-sensing and geospatial information, data from administrative sources, data from other harmonized business surveys, and data from the introduction of models will be used to derive the estimates of the 2021 CEAG. Section 3 presents an agile personalized collection instrument, while Section 4 shows some challenges and workarounds in implementing this project. Finally, Section 5 provides a brief conclusion, suggestions for future work and discussion questions for the panel.

# 2 The Canadian Census of Agriculture

## 2.1 Description

Agricultural data have been collected in Canada since 1666, and 2021 will mark the 23rd CEAG since Confederation.[2] The census paints a sweeping picture of the agricultural sector. It tracks changes in crops and livestock, as well as the evolution of farming practices and mechanization.

Statistics Canada has the legal obligation under the *Statistics Act* (R.S.C., 1985, c. S-19) to conduct the CEAG every five years. It provides a comprehensive picture of the agricultural sector at the national, provincial and subprovincial levels. The CEAG directly supports decision making and analysis at detailed geographical levels and is the only source of standard national information for small areas. It is used to support a number of important legislative requirements.

## 2.2 Use of alternative data sources in the Agriculture Statistics Program

Just like many programs at Statistics Canada, the Agriculture Statistics Program has been using an increasing amount of administrative data for statistical purposes over the last several years (Brackstone, 1987). Several factors contribute to the use of administrative data, including budget constraints, response burden reduction, an increased demand for estimates for small domains and more granular statistics, advances in technology that make processing large datasets more effective, and declining response rates. CEAG analysts use the administrative data for updates to the survey frame, data validation, edit and imputation, direct replacement of data, and estimation. Detailed information on revenues and expenditures in the 2016 CEAG content has been replaced in its entirety by data from the Agriculture Taxation Data Program (Hunsberger and O'Neill, 2016).

The Agriculture Division has long been one of the largest users of administrative data within Statistics Canada, with over 300 separate administrative datasets (both public and private data sources) among its holdings. A number of its programs are completely driven by administrative data. These data sources include sources from the Canada Revenue Agency (e.g., tax data) and supply-managed sectors (including dairy, chicken, eggs and turkey), where datasets include quota and production figures. Another data source is crop insurance agencies, which detail what crops have been planted and insured, as well as their yield at the field level.

2. Confederation: Canada became a country, the Dominion of Canada, in 1867.

The Agriculture Division also has a long history of employing satellite imagery in its programs, with the release of a crop condition assessment over two decades ago. This innovation has made Statistics Canada the first statistical agency to replace a survey by a satellite-imagery-driven model, with the replacement of the September Field Crop Reporting Series (Reichert et al., 2016).[3] These major datasets complement a range of diverse data obtained from administrative sources, including land valuations, grain marketing, goat-milking parlour registrations, honey bee permits, winery establishment grants and food processor data.

## 2.3 Alignment of concepts and definitions

As a first step, considerable efforts are being made to harmonize concepts and definitions within the Agriculture Statistics Program to facilitate the use of data between programs (validation and replacement). The Agriculture Statistics Program includes close to 40 surveys. Moreover, for the first time, the CEAG will be using Statistics Canada's business survey processing infrastructure, which will help harmonize methods, concepts and processes not only with the other agriculture surveys, but also with other economic statistics programs (nearly 200 surveys).

When the concepts or definitions from the CEAG are different than the ones from an alternative data source, one option is to work upstream in partnership with the data owner to see whether they can be harmonized. This improves the coherence and the relevance of the data, facilitates the file linkage, and reduces the cost. For example, there are many certifying bodies for organic food and farming in Canada, and the administrative data files they provide to Statistics Canada differ significantly (format, concepts, definition, coverage, quality, reporting period, etc.). The CEAG team is thinking about developing a collection and reporting tool in Microsoft Excel and providing it for free to the different certifying bodies. Another option is to change or modify the CEAG questions so that they better align with the administrative data sources.[4]

## 2.4 2021 Census (a hybrid approach)

The increasing availability of administrative data and satellite imagery provides Statistics Canada with the opportunity to eliminate all or almost all contact with agricultural producers by the 2026 CEAG. The objective is to reduce the response burden by 100,000 hours.

CEAG analysts are developing a new model that will be implemented by combining remote-sensing and geospatial information, administrative data, data from other harmonized business surveys, and data from the introduction of models. Other non-traditional alternative sources of information, such as web scraping or precision agriculture, will also be considered. This new model will be deployed progressively with the 2021 CEAG.[5] A proof of concept will be produced using the new model by predicting all the census variables (nearly 200 variables) for the whole population (close to 190,000 units). The CEAG concepts and definitions are being used as a framework for developing the CEAG-0 project.

3. Work is underway to replace the November Field Crop Reporting Series (produce estimates on harvested field crop areas, average yields and production) with alternative data sources and to develop a weekly yield model.

4. For example, the total number of employees receiving a wage or a salary reported in the 2016 CEAG might include employees that were not reported to the government (e.g., family members working unpaid hours). To better align the CEAG data with the administrative data, the question has been changed to ask about the total number of employees receiving a wage or salary as reported to the Canada Revenue Agency.

5. The key drivers in planning the 2021 CEAG are the following:

- reduce the agency's costs and the response burden put on agriculture operators while still maintaining a high-quality product
- maintain data coverage, relevance and quality
- reduce response burden by evaluating and adopting new data production and collection methods that go beyond the traditional survey approach by optimizing the use of existing functionalities in the electronic questionnaire
- further harmonize CEAG methods, concepts and processes with the economic statistics program to make analyzing and accessing integrated data easier for users, and to help them make informed, evidence-based decisions.

## 3 The new data collection strategy

### 3.1 An agile collection instrument—the inclusion of on-off switches in the collection instrument

Just like in 2016, the primary collection mode for the 2021 CEAG program will be the electronic questionnaire (EQ). Its functions will be maximized to reduce response burden by decreasing the time required to complete the questionnaire. The EQ will also improve data quality. For example, question flows will be optimized to limit the number of questions for respondents. This will be done by incorporating information coming from multiple sources (e.g., the Business Register) with an initial module in the EQ collection module regarding the type of activities (production) on the farm. This merger of information will ensure that only relevant questions are posed to the respondent. It is estimated that 70% of respondents will use the electronic collection mode, up 55% from the 2016 Census.[6] Paper questionnaires will be provided to respondents only upon request. Since some basic edits will be embedded in the EQ, fewer follow-ups with respondents are anticipated. This new collection instrument will be tested in May 2019, and adjustments will be made to the collection and follow-up systems thereafter.

When an alternative data source is of sufficient quality, the questions will be hidden using an agile collection instrument that allows data to be "smartly replaced."[7] The collection tool is built in such a way that the census team can choose whether a question will be asked directly to the respondent or whether an alternative data source will be used instead (i.e., the "on-off switches"). In essence, this would become a personalized EQ. Therefore, the number of questions to be answered will vary from one farm unit or respondent to another, and the final estimates for a specific variable might be the result of a hybrid of survey and administrative data. A shorter CEAG questionnaire will reduce the burden on respondents and should result in a lower number of incomplete questionnaires (partial response).

### 3.2 Smart replacement

To determine whether the quality of an alternative data source is good enough to potentially replace survey or census data, 2011 and 2016 CEAG estimates are being reproduced using these alternative data sources. A data confrontation exercise will then be performed at the micro and aggregate levels to measure the accuracy of these new sources. Any discrepancies will need to be explained (e.g., definition, concepts, reporting period and imputation), and, if they are deemed important and there is a business case, then a reconciliation process will take place. The challenge is to remove these differences from the microdata files or the calculation of the estimates.[8] Once the CEAG team has determined that the quality of an alternative data source is good enough and it can directly replace the survey or census data, the results are presented to the CEAG steering committee. The committee members either approve or reject the new approach. In the latter scenario, these estimates will still be used to either validate census data or impute in cases of non-response.

One of the biggest advantages is the fact that the model is personalized and works at the farm-unit level, question by question. The decision to use an administrative data source is made at this level. Therefore, the model is not restricted to administrative data files that have complete population coverage. The replacement model is flexible enough to allow the CEAG program to use the strengths of each data source.

Another advantage is that once the process to replace census data with alternative data has been established and approved, the estimation and validation processes can begin when the data are available. This will help to improve timeliness and disperse the analysis over a longer duration. Surprisingly, the approach developed using alternative data sources was found to sometimes provide better quality data. This is particularly the case when the concepts and definitions are not always understood by the respondents. The CEAG team also realized that the cost of

6. The option of completing the CEAG questionnaire online was first given in 2006. The Internet response rate was 5% at that time. It increased to 11% in 2011.
7. The working assumption is that up to 10 questions will be replaced in the 2021 CEAG by alternative data.
8. For example, the CEAG might directly use data coming from the "quota" files (e.g., dairy, chicken, eggs and turkey), but adjustments would be required to take into account the undercoverage.

producing the estimates for some variables is negligible once the approach is well established. Therefore, estimates for these variables could be released annually instead of every fifth year.

## 4 Challenges in moving forward with this new model

The biggest challenge CEAG-0 presents relates to the linkage of multiple data sources. In other words, how can CEAG-0 establish relationships between farm business units for which we have alternative data and farm business units to be surveyed from the Business Register (BR)? Luckily, the BR contains many variables that can be used to match records from different sources (business number, legal name, operating name, address, name of the farm operator, etc.). Most of the administrative data sources match relatively well, since they contain information about the same business unit or farm, but sometimes there are inconsistencies (false links or missed links). One of the advantages of the record linkage is that it increases the number of variables not found on the CEAG questionnaire, which help draw a more accurate picture of the agriculture industry and increase the relevance and usefulness of the census.[9] Note that Statistics Canada undertakes microdata linkages only in cases where the public good is clearly evident and outweighs the privacy intrusion. Moreover, the confidentiality of information relating to farm units used in microdata linkages is strictly maintained, and the results of the microdata linkage will not be used for purposes that can be detrimental to the farm units whose information is involved.

Administrative data have their own problems, since they are not necessarily collected for statistical purposes. Therefore, before administrative data can be used for the CEAG or any other surveys, they have to undergo a number of processing steps that transform them into data suitable for survey use or more appropriate for statistical use. Statistics Canada is working to develop strategies to overcome these shortcomings. For example, the Administrative Data Division (ADD) has been developing and maintaining databases of processed tax data and other databases of administrative data that have been used at Statistics Canada for more than 20 years. ADD is also helping the CEAG program to acquire administrative data by providing a corporate approach and common tools to rely on. Statistics Canada has also launched the Administrative Data Pre-Processing Project. The objective of this project is to create a corporate pre-processing service for administrative and alternative data with supporting solutions. This centralized corporate service will support all areas of Statistics Canada and will help reduce the need to create stand-alone solutions for a specific program. The project will implement the administrative data pre-processing framework with the ability to receive, capture, transform (format and standardize) and code (classify) administrative or alternative data records while consistently applying standard validation rules.

Last, but not least, the investment costs required to execute the CEAG-0 project are estimated at $“x” million. However, once completed, CEAG-0 could generate savings between $“y” million and $“z” million for each subsequent CEAG, producing a positive return on the investment of public funds.

## 5 Conclusion, future work and discussion questions for the group

A prudent approach has been adopted to implement the CEAG-0 to ensure that the quality of CEAG estimates are maintained. Efforts continue to be made to

- seek ways to maximize the use of administrative and alternative data sources to reduce response burden and ultimately reduce costs
- identify and evaluate the use of other potential administrative and alternative data sources from the public and private sectors
- develop partnerships with potential data providers
- enhance the processing and analysis of these sources to transform their data into data that are fit for survey or census use or more appropriate for statistical use

9. The CEAG links its information with the variables obtained from the Census of Population to produce a rich and detailed socioeconomic representation of the farm operators in the country at a detailed level of geography.

- improve the statistical methods that can be used to combine data from different sources (record linkage)
- harmonize the concepts and definitions from different data sources (i.e., improve data coherence)
- improve the overall quality of the CEAG estimates.

Here are a few discussion questions:

1. What are your experiences in the use of multi-source data? What are the main challenges that your organization has faced?
2. What are your experiences in measuring the quality of estimates from a hybrid of survey and administrative data sources?
3. What are your experiences and lessons learned in working with data providers or owners (e.g., building partnerships with them and collaborating in a mutually beneficial way)? Have you made any attempts to standardize and harmonize some concepts and definitions?

## 6 Acknowledgements

The author would like to thank Greg Peterson, Étienne Saint-Pierre and the Agriculture Division for all of their input and contributions to this project.

## References

Aston, C., F. Bédard, C. Mohl, G. Reichert, D.J. Valéry, B. Wesley and Z. Yinsuo. (2016). Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Statistical Survey Data. Paper presented at ICAS VII Seventh International Conference on Agricultural Statistics, available on: https://www.istat.it/storage/icas2016/g43-reichert.pdf.

Brackstone, G.J. (1987). Issues in the Use of Administrative Records for Statistical Purposes. Survey Methodology, 13, 29-43.

Hunsberger, P., and L. O'Neill. (2016). Potential uses of tax data in the Canadian census of agriculture. Paper presented at ICAS VII Seventh International Conference on Agricultural Statistics, available on: https://www.istat.it/storage/icas2016/pl2-hunsberger.pdf.

Michaud, S. (2017). Modernization at Statistics Canada. Presentation made at the 4th International Conference on Big Data for Official Statistics, organized by the UN Global Working Group, DANE and the Colombian Ministry for ICT, available on: https://unstats.un.org/unsd/bigdata/conferences/2017/presentations/day1/session1/3 - Sylvia Michaud - Statistics Canada.pdf.

Statistics Canada. (2016). Questionnaire for the Canadian Census of Agriculture, available on: http://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=getInstrumentList&Item_Id=235427&UL=1V.